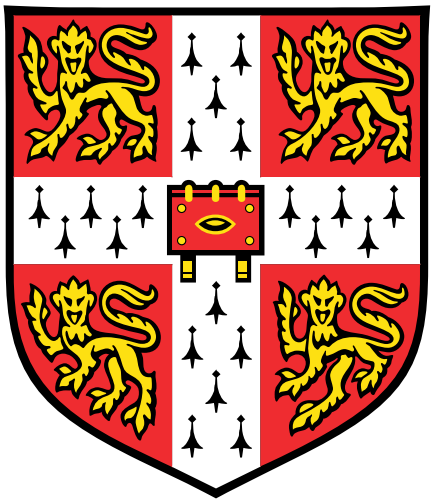
Digital phenotyping through multimodal, unobtrusive sensing



Ignacio Perez-Pozuelo

Jesus College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

August 2020

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared and specified in the text preceding each chapter. It is not substantially the same as any that I have submitted, or is being concurrently submitted, for a degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution.

As required by the Degree Committee for the Faculties of Clinical Medicine and Veterinary Medicine, this dissertation contains fewer than 60,000 words.

Ignacio Perez-Pozuelo
August 2020

ABSTRACT

Digital phenotyping through multimodal, unobtrusive sensing

Ignacio Perez-Pozuelo

The growing adoption of multimodal wearable and mobile devices, such as smartphones and wrist-worn watches has generated an increase in the collection of physiological and behavioural data at scale. This digital phenotyping data enables researchers to make inferences regarding users' physical and mental health at scale, for the first time. However, translating this data into actionable insights requires computational approaches that turn unlabelled, multimodal time-series sensor data into validated measures that can be interpreted at scale.

This thesis describes the derivation of novel computational methods that leverage digital phenotyping data from wearable devices in large-scale populations to infer physical behaviours. These methods combine insights from signal processing, data mining and machine learning alongside domain knowledge in physical activity and sleep epidemiology. First, the inference of sleeping windows in free-living conditions through a heart rate sensing approach is explored. This algorithm is particularly valuable in the absence of ground truth or sleep diaries given its simplicity, adaptability and capacity for personalization. I then explore multistage sleep classification through combined movement and cardiac wearable sensing and machine learning. Further, I demonstrate that postural changes detected through wrist accelerometers can inform habitual behaviours and are valuable complements to traditional, intensity-based physical activity metrics. I then leverage the concomitant responses of heart rate to physical activity that can be captured through multimodal wearable sensors through a self-supervised training task. The resulting embeddings from this task are shown to be useful for the downstream classification of demographic factors, BMI, energy expenditure and cardiorespiratory fitness. Finally, I describe a deep learning model for the adaptive inference of cardiorespiratory fitness (VO_2max) using wearable data in free living conditions. I demonstrate the robustness of the model in a large UK population and show the models' adaptability by evaluating its performance in a subset of the population with repeated measures 6 years after the original recordings.

Abstract

Together, this work increases the potential of multimodal wearable and mobile sensors for physical activity and behavioural inferences in population studies. In particular, this thesis showcases the potential of using wearable devices to make valuable physical activity, sleep and fitness inferences in large cohort studies. Given the nature of the data collected and the fact that most of this data is currently generated by commercial providers and not research institutes, laying the foundations for responsible data governance and ethical use of these technologies will be critical to building trust and enabling the development of the field of digital phenotyping.

ACKNOWLEDGEMENTS

I am deeply grateful to Professor Cecilia Mascolo for guiding me through this PhD and for her encouragement, support and intellectual curiosity. It is thanks to her and the freedom I was given that I was able to focus on areas of research that I was truly passionate about and to explore multiple areas of digital health. I would also like to thank Dimitris Spathis, who became Amos to my Daniel, I couldn't have asked for a better partner in research. The period between August of 2019 and March of 2020 was truly one of the most intellectually stimulating periods of my life and I hope that we can continue working together for years to come. Similarly, I am very thankful to João Palotti, Luis Fernandez Luque, Marius Posa, Bing Zhai, Tom White, Ian Tang and Josh Cowls who helped shape my research and taught me a lot throughout the process. I would like to thank my coauthors in Cambridge, Oxford, MIT, Newcastle, The Alan Turing Institute, QCRI and Weill Cornell for all their support during my PhD. I would also like to thank Professor Edwin Robertson for his mentorship, support and understanding through all of my graduate studies. Many thanks to Soren Brage and the rest of the PA team for helping me as much as they could and for recruiting me to Cambridge. My thanks also to Professor Nick Wareham who, while leading one of the world top institutes, always found time to give me valuable advice in research and life. Finally, I would like to thank GlaxoSmithKline and the Engineering and Physical Sciences Research Council for supporting my PhD and the Alan Turing Institute for a fantastic Enrichment Year.

Pursuing a PhD can be really tough and I cannot imagine what I would have done without the support of loyal friends who were going through the same ups and downs. My social life as a PhD student at Cambridge has been largely enabled by friends I made through rowing, and I am very thankful to my friends at CUBC, Cambridge 99, Club Nautico de Sevilla and Jesus College Boat Club. Similarly, my time at lab was enriched by group lunches at the Clinical School and later by the incredibly welcoming Mobile Systems Group in the Computer Lab. Finally, I am blessed to have an incredible network of friends who are now doing outstanding things all around the world from both coasts of the USA, to Asia and other parts of Europe who have also been a great source of support throughout this process.

I want to thank my family for their unfailing love and support for me. More than words could ever express, I am grateful for Emma and her love and support (even at the craziest of times) throughout this process. Emma's help has been immense both from a personal standpoint and

in my own academic work through thorough proofreading and fierce intellectual discussions. Finally, I would like to dedicate this thesis to my parents and sister, who have supported me through the good and the bad for many years now and to Dr. Angel Esteban and Dr. Julio Prieto who kickstarted my interest in research at a very young age.

PUBLICATIONS AND PRESENTATIONS

Publications resulting from this thesis

Journal publications

Perez-Pozuelo, I. White, T., Westgate, K., Wijndaele, K., Wareham, N. J., & Brage, S. (2019). Diurnal profiles of physical activity and postures derived from wrist-worn accelerometry in UK adults. *Journal for the Measurement of Physical Behaviour*, 1(aop), 1-11.

Perez-Pozuelo, I., Zhai, B., Palotti, J., Mall, R., Aupetit, M., Garcia-Gomez, J. M., ... & Fernandez-Luque, L. (2020). The future of sleep health: a data-driven revolution in sleep science and medicine. NPJ digital medicine, 3(1), 1-15.

Zhai, B.*, **Perez-Pozuelo, I.***, Clifton, E. A., Palotti, J., Guan, Y. (2020). Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(2), 1-33.

Perez-Pozuelo, I., Spathis, D., Gifford-Moore, J., Morley, J., Cowls, J. (2021). Digital phenotyping and sensitive health data: Implications for data governance. Journal of the American Medical Informatics Association.

Tang, C. I., **Perez-Pozuelo, I.***, Spathis, D.*, Brage, S., Wareham, N., Mascolo, C. (2021). SelfHAR: Improving Human Activity Recognition through Self-training with Unlabeled Data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies

Conference papers

Zhai B.*, **Perez-Pozuelo I.***, Brage S., Guan Y. (2019). Ubiquitous monitoring of sleep-wake cycles using combined sensing and deep learning models. *In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE.*.

Zhai, B.*, **Perez-Pozuelo, I.***, Clifton, E. A., Palotti, J., Guan, Y. (2020). Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. UBICOMP 2020, UBICOMP 2021.

Tang, C. I., **Perez-Pozuelo, I.***, Spathis, D.*, Brage, S., Wareham, N., Mascolo, C. (2021). SelfHAR: Improving Human Activity Recognition through Self-training with Unlabeled Data. UBICOMP 2021

Spathis, D., **Perez-Pozuelo, I.**, Brage, S., Wareham, N., Mascolo, C. (2020) Learning Generalizable Physiological Representations from Large-scale Wearable Data. NIPS (mHealth).

Tang, I., Spathis, D., **Perez-Pozuelo, I.**, Mascolo, C. (2020) Exploring Contrastive Learning in Human Activity Recognition for Healthcare. NIPS (mHealth).

Spathis, D., **Perez-Pozuelo, I.**, Brage, S., Wareham, N. J., Mascolo, C. (2021). Self-supervised transfer learning of physiological representations from free-living wearable data. ACM Conference on Health, Inference, and Learning (ACM-CHIL).

In preparation or under review

Gonzales, T. I.*, Jeon, J. Y.*, Lindsay, T., Westgate, K., **Perez-Pozuelo**, I., Hollidge, S., Brage S. Wareham, N. (2020). Resting heart rate as a biomarker for tracking change in cardiorespiratory fitness of UK adults: The Fenland Study. Heart.

Perez-Pozuelo, I., Posa, M., Spathis, D., Westgate, K., Wareham, N., Mascolo, C., Brage S. Palotti, J. (2020). Detecting sleep in free-living conditions without sleep-diaries: a device-agnostic, wearable heart rate sensing approach. Nature Digital Medicine.

Palotti, J.*, Posa, M.*, **Perez-Pozuelo, I.*** (2021). HypnosPy: A Device-Agnostic, Open-Source Python Software for Wearable Circadian Rhythm and Sleep Analysis and Visualization. Sensors (Updating camera-ready version).

Hasthanasombat, A., Spathis, D.*, **Perez-Pozuelo, I.***, Mascolo, C. (2020) Exploring generalisability in human activity recognition through minimising invariant risk. NIPS (ML4H).

Other Publications

Book Chapters

Perez-Pozuelo, I., Spathis, D., Clifton, E. A., Mascolo, C. (2020). Wearables, smartphones, and artificial intelligence for digital phenotyping and health. Digital Health: Mobile and Wearable Devices for Participatory Health Applications. Elsevier.

Conference presentations

Past, Present and Future of AI in Sleep (2021). American Academy of Sleep Medicine (AASM) (Virtual) *Oral presentation, panelist*

Making Sense of Sleep (2020,2021). UBICOMP (Virtual, TBD)

Sleep epidemiology. (2019). IEEE Engineering in Medicine and Biology Conference (EMBC). Berlin, Germany. *Oral presentation*

Sleep-wake classification using multimodal sensing. (2019). IEEE Engineering in Medicine and Biology Conference (EMBC). Berlin, Germany. *Workshop*

* Equal contribution to this work

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PRINCIPAL ABBREVIATIONS

Symbols

$\mathscr{D}$	Dataset
$\mathscr{L}$	Loss function
W	Weight matrix
X	$\mathscr{D}$ inputs (matrix with N rows, one for each data point)
Y	$\mathscr{D}$ outputs (matrix with N rows, one for each data point)

Acronyms and abbreviations

AASM	American Academy of Sleep Medicine
BMI	Body mass index
CI	Confidence interval
CNN	Convolutional neural network
DXA	Dual energy x-ray absorptiometry
ECDF	Empirical Cumulative Distribution
ENMO	Eucledian Norm Minus One
GRS	Genetic risk score
GWAS	Genome-wide association study
LSTM	Long-short term memory
MESA	Multiethnic study of atherosclerosis
MET	Metabolic Equivalent
MVPA	Moderate to vigorous physical activity
NREM	Non-rapid eye movement

Principal abbreviations

PC	Principal component
PPG	Photoplethysmography
PSG	Polysomnography
RCT	Randomised controlled trial
REM	Rapid eye movement
RHR	Resting Heart Rate
RNN	Recurrent neural network
SD	Standard deviation
SDS	Standard deviation score
SE	Standard error
SSA	Singular spectrum analysis
t-SNE	t-Distributed Stochastic Neighbor Embedding
UKB	UK Biobank
VPA	Vigorous physical activity
WHO	World Health Organisation

CHAPTER 1

INTRODUCTION: WEARABLE AND MOBILE DEVICES FOR DIGITAL PHENOTYPING AND HEALTH

Ubiquitous progress in wearable sensing and mobile-computing technologies, alongside growing diversity in sensor modalities, has created new pathways for the collection of health and well-being data outside of laboratory settings, in a longitudinal fashion. Traditionally, epidemiologists and clinicians have relied upon self-report measures of physical activity and sleep which, whilst valuable in the absence of alternatives for large-scale data collection, are subject to bias and often provide partial, incomplete information [2]. Wearable and mobile devices now have the potential to provide low-cost, objective measures of physical activity, clinically relevant data for patient assessment and scalable behaviour monitoring in large populations. Indeed, today, over 400,000 participants have had their behaviour tracked prospectively using accelerometers for epidemiological studies across the globe [3–5]. This data can be used in both interventional and observational studies to derive insights regarding the links between behaviour, health and disease, as well as to advance the personalization and effectiveness of commercial wellness applications.

Physical behaviour data extracted from wearable devices is now being used to derive sensor-assessed, objective measures of physical behaviours, overcoming the limitations of self-report with the aim of relating these to clinical endpoints and eventually applying the findings to preventive and predictive medicine. Moreover, the application of artificial intelligence (AI), sensor fusion and signal processing to wearable sensor data has led to improved human activity recognition (HAR) and behavioural phenotyping. In this chapter, we first introduce the state of the art in wearable and mobile sensing technology in epidemiology and clinical medicine and discuss how AI is changing the field. We then discuss the outstanding challenges in HAR and the inference of sleep and fitness from wearable sensor data, alongside the contribution of the original research work presented in this thesis to addressing them.

1.1 Towards Digital Phenotyping

Physical behaviours, including activity and sleep, have long been thought to have important health consequences. The earliest records of organised exercise as a formal means of health promotion date back to 2500 BC [6] and the importance of physical activity and exercise to health and longevity have long been promulgated. For instance, Hippocrates and Galen advised that a lack of physical activity, as well as over-exertion, were detrimental to health and well-being. Whilst the links between physical behaviour and health have long been suspected, the formal, objective investigation of this relationship began with the *Industrial Revolution* in England. The first physical activity studies were small, cross-sectional, observational studies, including those conducted by Dr WA Guy of King's College London, who compared mortality rates between physically active and sedentary workers, favouring the former [7]. Modern-day physical activity and exercise science research may have originated during the mid-1800s due to the concern for the well-being and longevity of rowing oarsmen from the Universities of Oxford and Cambridge [8]. Until then, the common belief, dating back to Galen, suggested that vigorous exercise had deleterious effects on health. However, these studies showed that the life expectancy of the athletes exceeded that of the general population, suggesting that even vigorous exercise might possibly be advantageous to health [9]. Whilst these early studies were all subject to potential biases and confounding, they sparked interest in understanding the relationship between activity and health outcomes, which persists to the present day.

The field of physical activity and health research as we now understand it developed during the 1950s with studies of the association between physical activity and health [10, 11]. Seminal work on coronary heart disease and physical activity amongst London workers by Morris and colleges was followed by similar studies in the US. During the 1980s, the volume of evidence suggesting and inverse association between physical activity and coronary heart disease, supported by reviews of both cross-sectional and longitudinal studies, afforded epidemiologists more confidence in asserting that the association was causal in nature [12]. The findings had a strong impact on the promotion of physical activity as a matter of urgent public health interest, ultimately leading to the physical activity recommendations launched in 1995 by the Centres of Disease Control and Prevention and the American College of Sports Medicine, alongside other analogous institutions elsewhere [13]. In the following years, randomised trials of physical activity interventions were conducted, providing further support for a causal impact of physical activity on health and the notion that physicians might prescribe physical activity began to become more prevalent in the literature [14]. By the turn of the millennia, alongside a growing number of epidemiological studies exploring the association between physical activity and health, studies had begun to explore the measurement of physical activity [15].

Until recently, the study of physical activity has been hindered by the inability to accurately quantify its component parts. The first large population-based study with objective monitoring of physical activity levels was presented in 2008 [16]. Importantly, technological advances

in wearable devices and smartphones increasingly facilitate the collection of vast amounts of multimodal data in an unobtrusive, seamless way. In particular, the use of data generated passively by these devices enables the measurement of free-living human behaviour in a scalable manner. This data can be used for digital phenotyping. Given evidence of the health consequences of physical behaviours, being able to distinguish exactly what, when and how physical activity is linked to health, through more accurate characterization of physical activity, would be advantageous.

Digital phenotyping can be defined as "movement-by-movement quantification of the *in situ* individual-level human phenotype using data from personal digital devices" [17]. This new field has already generated significant research interest across epidemiology and clinical medicine. For instance, in psychiatry, objective, multimodal, continuous quantification of behaviour using individuals' own devices may result in clinically useful markers which can then be used to improve diagnostics, tailor treatment or design new intervention models [18]. Similarly, real-time feedback paired with AI models introduces new opportunities for health and well-being applications. For example, it may become possible to develop personalised interventional feedback generated automatically based upon physiological, environmental and social cues from mobile and wearable devices [19].

The decreasing cost and increasing capabilities of sensors embedded in mobile and wearable devices, coupled with the proliferation of data sources from social media, environmental sensors and other sources have yielded new ideas and techniques in the study and quantification of well-being, mobility and social interaction [19]. However, as the data is proliferating, there remain important limitations to our ability to interpret this data, allowing us to make inferences about the associations between physical behaviour, health and disease.

Throughout this thesis we leverage multimodal wearable sensor technologies from large epidemiological cohort studies, with strong parallels to the data captured by wearable devices, to derive inferences about physical behaviours and characteristics. For instance, we explore how these signals can be used to determine an individuals' cardiorespiratory fitness (CRF) by inferring their VO_2max or to determine their sleep timing and quality. Improving our ability to infer these characteristics from the data, helps to address an important limitation of current research efforts.

Remark: The remainder of this chapter provides an introduction to how multimodal wearable and smartphone devices can be used to derive objective measurements of physical activity and behaviour. In doing so, we provide an introduction to the field of physical activity epidemiology and the transition from questionnaire based assessments to objective monitoring through accelerometers. We explore how mobile phones can be used to track physical and psychological behaviours. Furthermore, the impact of AI in this emerging field of digital phenotyping is explored and the role of this thesis in advancing research in this field is addressed.

1.2 Objective measures of physical behaviours in epidemiology

1.2.1 Introduction to epidemiological research

The overarching goal of epidemiological research is to inform the development of interventions that reduce mortality and morbidity in populations [20]. In order to achieve this aim, epidemiologists study the distribution of health-related states or events, such as disease, in order to understand their burden and identify their determinants [1]. To conduct this type of research, intersecting data regarding both the outcome of interest and the potential determinants is required. Not only must this data intersect, with the same individuals providing information about both the exposure and the outcome, but it must be both reliable and valid [20]. This means that the measure used to assess the exposure and the outcome must be repeatable over time and accurately convey what it intends to measure. In general, objective measures are preferred, whereby individuals are not required to recall or report their exposure or outcome status themselves. This protects against unintentional recall biases and inaccuracies, as well as intentional adjustments to reporting based on social desirability. However, these considerations must also be balanced against the burden objective measuring places upon participants and the other costs that they incur. Researchers may favour a marginally less accurate measure if the measure can be collected with ease at low cost and is less unlikely to be refused by participants, such that many participants can be included, thus increasing the power of the study to detect associations [20].

In planning studies and drawing conclusions, epidemiologists must be attentive to chance, bias, confounding and reverse-causality that could cause them to draw erroneous conclusions. For example, if a study reported an association between sleep duration and obesity, the results must be interpreted with caution and cannot be assumed to constitute evidence of a causal relationship without further criteria being met [20]. In this example, the relationship could be spurious and simply the result of chance. The probability of chance explaining the results diminishes as the number of studies reporting the same finding increases. Further, the probability of chance diminishes if larger data-sets are used. If the result is not spurious, it may be the result of reverse causality. Contrary to the initial hypothesis, obesity may be the exposure variable and sleep may be the outcome. Various methods to help rule-out reverse causality exist. At minimum, longitudinal studies are required such that sleep measures are collected prior to the onset of obesity. Further, if exposures are amenable, Randomised controlled trials (RCTs) can be conducted or, if the genetic determinants of an exposure are well characterised, Mendelian randomisation (MR) analyses can be performed. For example, a 2019 MR study using genetically predicted sleep duration as the exposure and BMI as the outcome, suggested a causal inverse association between sleep duration and BMI in children [21].

[1] https://www.who.int/topics/epidemiology/en/

If reverse causality does not appear a likely explanation it remains possible that confounding from a third, extraneous variable, associated with both the exposure and the outcome but which does not lie on the causal pathway between them explains the relationship [20]. For instance, smoking may cause both poor sleep and obesity, inducing a statistical association between the two variables that may erroneously be interpreted as a causal relationship. In order to account for confounding, analyses should be controlled for potential confounders or MR analyses using genetic instruments may be used.

Finally, various forms of bias should be considered. Together these comprise systematic errors in the design, conduct or analysis of a study that may result in a distortion of the relationship between exposures and outcomes [22]. There are two major sources of bias in epidemiological research: selection bias and information bias [20]. Selection bias relates to the study population in which a research question is addressed. For example, if a study includes only adult men, the results are only generalizable to adult men and cannot be considered applicable to other population groups. A common source of selection bias in epidemiological research is that the individuals who choose enrol in population-based studies are often healthier and better educated than the general population from which they are drawn [23, 24]. Whilst the results of studies that suffer from selection bias are still internally valid, they must be generalized with caution. Biases may also relate to the data collected. This is referred to as information bias and will be elaborated in the following section.

Overall, epidemiological research requires large data-sets with accurate, cost-effective and minimally burdensome measures of exposures, outcomes and potential confounding variables. Ideally, these data-sets should follow participants longitudinally. In the following sections, the way in which multimodal wearable sensing devices have revolutionised the ability to interrogate the associations of human activity to health and disease is explored. These abilities increasingly facilitate accurate epidemiological research at scale in this area for the first time.

1.2.2 Traditional measurement of physical activity through questionnaires

Prior to the advent of wearable sensing and mHealth technologies, researchers primarily relied upon questionnaire-based methods to measure physical activity. Questionnaires have many advantages for epidemiological research. They do not require experts or any special training to administer, they are also cost-effective, non-invasive and widely acceptable to participants. Further, individuals can be asked to report upon their typical, long-term habits and behaviours which may not be accurately represented in laboratory settings. These characteristics of questionnaires facilitate the collection of data from large numbers of individuals and explains the popularity of these approaches [2].

[2] https://www.who.int/ncds/surveillance/steps/resources/GPAQ_Analysis_Guide.pdf

Despite their many advantages, as alluded to in the previous section, questionnaires are not objective measures and may be subject to information bias. Information bias occurs when the measures used in a study are inaccurate. In the case of self-report measures, individuals may inaccurately recall their behaviour, report an idealised version of their habits or some combination. Previous studies have found that self-reported physical activity suffers from reporting bias and that this results from a combination of social desirability bias (reporting behaviour which is seen to be socially desirable), as well as the cognitive complexity of reporting the duration, intensity and frequency of physical activity behaviours with precision [16, 25, 26]. In addition, the understanding of a behaviour that is self-reported is limited to the specific set of questions given to study participants. These may not be enough to reflect a complete view of complex behaviours. Inaccuracies resulting from reporting errors may be randomly distributed across the population being studied. In this case, the results of the study would be biased toward the null, diminishing the ability of the researchers to identify true associations between exposures and outcomes. However, the errors may also be systematic, with participants in different population groups systematically under or over-reporting their activity levels. This could lead to associations being identified which are the result of bias, rather than cause, and leading to the drawing of erroneous conclusions. This could have serious implications if public health guidance were to be based on these conclusions.

In order to diminish concern regarding information bias in studies using self-report measures of physical activity, questionnaires should ideally be validated against a gold-standard measure.

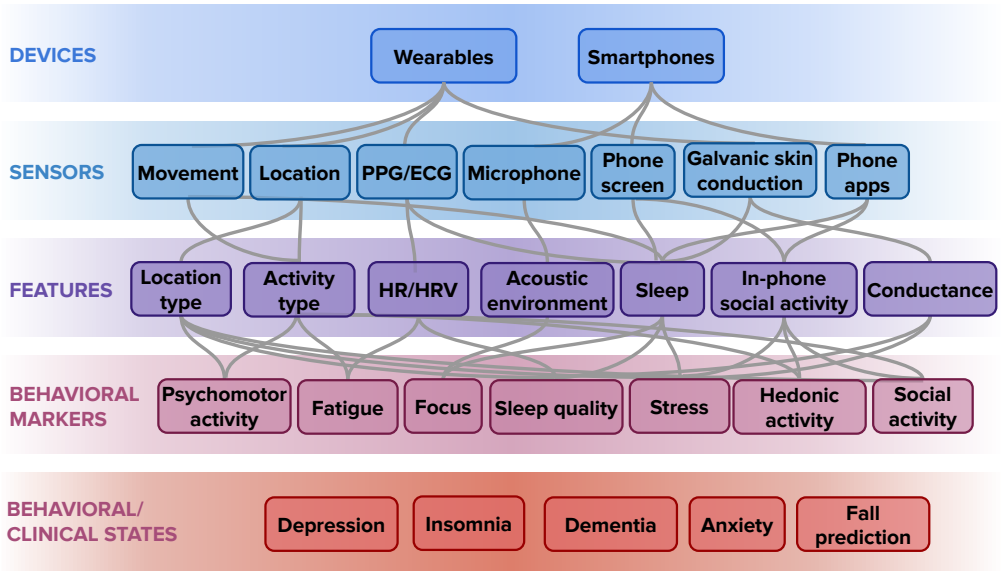
1.2.3 The transition towards objective monitoring of physical behaviours

Objective monitoring of physical activity with devices such as pedometers (step measurement [27–29]), actigraphy (count-based movement measurement [30]) and accelerometers (raw movement intensity measurement) have been used to overcome the limitations of self-reported activity measures [31]. Recently, increasingly sophisticated sensors embedded within smartphones have resulted in a proliferation of *affective computing* and *behavioural phenotyping* applications. A non-comprehensive overview of the current landscape for human behaviour phenotyping using wearable sensors and smartphones is presented in Figure 1.1.

Technological advances in the last 20 years allow for devices like triaxial accelerometers to record and store data across multiple days without requiring recharging. Further, such devices are increasingly affordable, reliable and non-obtrusive. Indeed, in 2003, the National Institutes of Health and the National Cancer Institute funded the National Health and Nutrition Examination Survey (NHANES) [3], a large epidemiological study that aims to understand the objective measurement of physical activity through accelerometery, became the first study of its kind in the United States. Many other large initiatives followed. The UK Biobank Study

[3] https://www.cdc.gov/nchs/nhanes/index.htm

Figure 1.1 Overview of the layered, hierarchical framework of mobile and wearable sensing technologies for digital phenotyping. The framework starts with the input layers where mobile and wearable devices yield different, constantly evolving sensing capabilities. The middle layers of the framework relate to inferences, from features to behavioural markers. The output layer comprises the delivery of these inferences and finally behavioural/clinical states identified through this framework (PPG: photoplethysmography, ECG: Electrocardiogram, HR: heart rate, HRV: heart rate variability). Figure inspired by Mohr et al., [1].



[4], the Whitehall study [5] and the China Kadoorie Biobank (CKB) [6] all exemplify the use of accelerometry in large-scale observational studies.

These studies allow researchers to perform epidemiological investigations exploring the associations between activity-related exposures of interest (predominantly comprising physical activity, sedentary behaviour and sleep) and disease outcomes, whilst controlling for potential confounders, such as diet, alcohol consumption, smoking habits or socio-economic status, also collected amongst the study participants. In addition, such large-scale studies often provide intersecting genome-wide genotyping information. This facilitates MR studies, as well as novel genome-wide association studies (GWAS) designed to identify the determinants of physical activity, sedentary behaviour or sleep [32]. These GWAS results can then be used to facilitate MR studies in other cohorts, designed to assess the causal impact of physical behaviours on health and disease outcomes [21].

[4] https://www.ukbiobank.ac.uk/

[5] https://www.ucl.ac.uk/epidemiology-health-care/research/epidemiology-and-public-health/research/whitehall-ii

[6] https://www.ckbiobank.org/site/

1.2.4 Analysing physical activity: Accelerometers for movement analysis

Although physical activity and exercise are often used interchangeably in the literature, there is a difference between these concepts. Physical activity can be defined as any bodily movement that results in *energy expenditure* being increased above resting levels. Exercise is a particular type of physical activity that is purposeful, planned, structured and often repetitive [33]. As such, activities such as housework are considered examples of physical activity, but not of exercise, because they are typically sporadic and unplanned in nature [34].

Physical activity can be broken down and defined by (1) type (walking, running, cycling, etc); (2) duration or volume (total time performing the activity); (3) frequency (number of sessions either per day or per week) and (4) intensity (how much energy is expended during exercise) [35]. Metabolic equivalent tasks (METs) are often used to describe the intensity of a given activity. For instance, one MET is equivalent to sitting at rest [35]. Depending on their intensity, activities can be categorized into: sedentary ($\leq$1.5 METs), light (1.6-2.9 METs), moderate (3.0-5.9 METs) or vigorous ($\geq$ 6.0 METs) [35]. Different types of activities will normally fall into one of these buckets repeatedly. For instance, typing on a computer would be categorised as sedentary, walking is considered light, brisk walking is moderate and running is vigorous. In order to understand physical behaviours at a population level, it is imperative to be able to accurately quantify the intensity of activities and link this to health outcomes. This informs physical activity recommendations, as well as the assessment of whether these recommendations are being met [34].

Accelerometry is a valuable technique for the accurate estimation of daily energy expenditure in large population studies, given its feasibility, low cost and the existence of validation studies. Acceleration signals are composed of a movement component, a gravitational component and noise [36]. When conditions are static with non-rotational movement, the gravitational component is visible as the offset of one or more sensor axes and can be used for the detection of the sensor orientation in relation to the vertical plane [36]. However, this separation is complicated when rotational movements are included as the frequency domains of the movement-related component and the gravitational component can overlap, making it almost impossible to separate these two components using simple frequency-based filtering [37]. The inclusion of gyroscopes in addition to accelerometry helps to mitigate this problem but they are not yet feasible for use in large-scale observational research [38, 36]. A schematic of the processing and analysis of raw accelerometer signals is presented in Figure 1.2. This process starts with raw measurements and data storage of triaxial acceleration waveforms (usually between 60-100 Hz), followed by a *post-processing step* where the sensor is calibrated to local gravity, time-stamping and re-sampling take place. The filtering of machine noise ($\geq$ 20Hz) follows and non-wear time is then identified. Once this *post-processing step* finishes, summary metrics and feature extraction follows. In this step, statistical metrics and features (i.e. mean magnitude, pitch, roll, power spectra, etc) are derived (see Fig. 1.2).

Several accelerometer-derived metrics and constructs are well-defined, established methods to quantify objective physical activity records. These metrics can then be used to then estimate energy expenditure (EE) and derive metabolic equivalents (METs) [39]. Some of the best established metrics are:

Volume of Physical Activity: Volume of physical activity refers to the total volume of activity in a given time period. In order to compare different records and recordings of different lengths, volume of physical activity is divided by the duration of the measurement to result in an average activity intensity rate.

Intensity: As previously mentioned, physical activity intensity can be categorized into: Vigorous, Moderate, Light and Sedentary. These categories were originally defined by asking participants but have since been informed by objective data, cross-referencing with resources such as the Ainsworth Compendium [40], which is an aggregation of mean activity intensities that are measured or estimated while performing different activities.

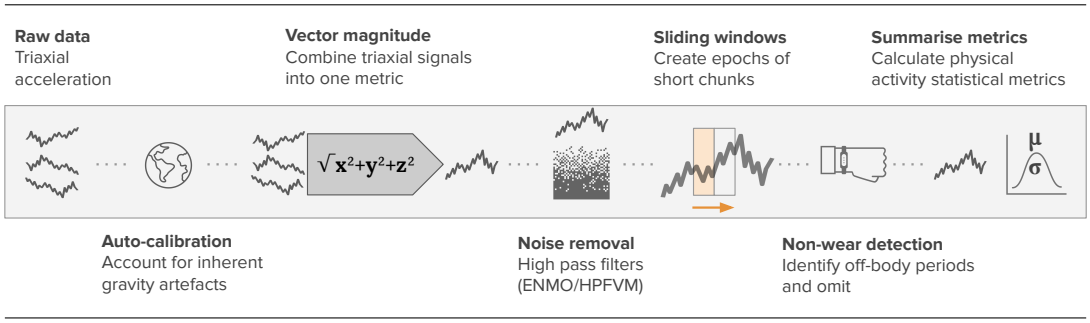
Posture: Posture, limb positioning and the pose of the body are of interest to physical activity scientists as they can provide new context for other measurements of physical activity, including work developed for this thesis [41, 42]. Indeed, interest in this domain has grown in recent years given the complementary nature of these inferences to traditional, intensity-based metrics also obtained through accelerometers. For instance, the consensus statement regarding the definition of sedentary behaviour now includes the sitting posture as a defining characteristic [43]. Advances in micro-electro-mechanical sensors (MEMs) and orientation estimation algorithms allow wearable sensors to be used for non-restricted human motion capture applications [44]. Bio-mechanically, human bodies are composed of a series of connected, jointed links that move and operate with different degrees of freedom (DOF) which can be measured using these devices. However, proper estimation of consistent and clinically meaningful joint kinematics using wearable inertial sensors requires a sensor-to-segment coordinate system calibration and understanding. To describe limb location, six parameters are required. These comprise: location ((x,y,z) coordinates with respect to the reference system axes) and orientation parameters ((α,β,γ) angles with respect to the reference system plane) of a limb segment in space. These six coordinates constitute the DOFs of a given limb segment in space and can be used to define orientation and spatial location at a given time.

1.3 Digital phenotyping in large population studies

1.3.1 Multimodal sensing

Most conventional studies using either smartphones or wearable sensors to study physical behaviours have used single sensor approaches for measurement and classification tasks (accelerometer, pedometers or gyroscopes) [3, 45]. Occasionally they have used GPS for

Figure 1.2 Typical data analysis pipeline for movement sensor data: from raw accelerometer data to appropriate filters and summary statistics. (*ENMO*: Euclidean Norm Minus One, *HPFVM*: High-pass Filtered Vector Magnitude)



coarse grained location sensing. However, smartphones and new generation wearable devices often come equipped with a vast array of sensors that enable multimodal sensing. Incorporating multimodal sensing information can yield additional physiological and environmental cues, such as sound, heart rate, skin conductance, location or activity type. Indeed, large-scale longitudinal studies, such as 'All of US' [7], will incorporate multimodal wearable sensor data with the aim of better understanding physical behaviours in *free-living* environments.

Multimodal sensing approaches often rely on traditional shallow models, like Random Forests or Support Vector Machines, operating on features extracted from each sensor separately [46]. Subsequently, there are two strategies to perform sensor fusion: *Feature Concatenation* ([47], [48]) that produces a single feature vector merging all the features extracted upstream; and *Ensemble Classifiers* ([49]) where classifiers are trained in single modalities and their predictions merged at the final step.

A significant challenge arises when attempting to incorporate information from sensor types that are different in nature (i.e. an accelerometer, an ECG, and a phone camera). Due to the inherent differences in sampling rates and data distributions or shapes, the aforementioned approaches struggle to merge these diverse inputs and produce meaningful representations. An important insight here is to combine and find patterns regarding the latent cross-sensor interactions that cannot be discovered in isolation or ensembles. This is achieved with shared or merged layers in deep neural networks that can model different sensor time-series and extra participant metadata in a joint latent space (see Fig. 1.6). However, this area remains an important area for improvement in the existing literature. Contributing to the addressing of this gap in the literature, the merging of accelerometry and heart rate sensor data is one of the topics addressed in this thesis.

[7] https://allofus.nih.gov/

1.3.2 A brief introduction to deep and representation learning

Data-driven insights derived from AI have already had a tangible impact on wearable sensing and mobile health. These developments have facilitated better HAR models, more accurate predictive models of human behaviour and the development of personalised lifestyle recommendations. In this section, two schools of thought are presented regarding the application of AI methods to wearable and mobile data. The first looks for informative features that represent the time-series through inventive feature extraction, whilst the second is based on the emerging power of representation learning to automatically extract features from lightly processed time-series during the training process.

1.3.2.1 Traditional feature-engineering modelling

Usually, mobile and wearable sensor data is transformed into *feature vectors* in order to ensure compatibility with the majority of machine learning algorithms. A feature vector is a matrix-like data structure where each row represents a unique sample and each column is a separate feature or variable. However, the raw time-series signals arising from, for example accelerometers, are represented as multiple continuous sequences. Consequently, the next step after data collection is to summarize the information from each sensor into a number of independent variables that capture semantic information. This task is called feature-extraction and researchers work to come up with increasingly complex features that correlate with a given label.

Depending on the size of the datasets and the computing power available, computing these features as a pre-processing step can be a time-consuming, multi-step process. Simple statistics such as the mean, median, standard deviation and inter-quartile range are easier to estimate and could be used. However, they may not capture the informative features of noisy signals. On the other hand, higher-order statistics and transformations like the kurtosis, skewness, stationary, least squares slope, autocorrelation, Fourier transform, and entropy provide more expressive metrics that reflect real time-series phenomena like the seasonality or repeatability [50, 51].

After the calculation of the appropriate metrics from the time-series signal, they are then fed to machine learning algorithms. If additional linked datasets (metadata) exist (i.e. demographic, anthropometric or personality traits) they are concatenated with the sensor features into a big feature vector. The most common classification algorithms found in the literature are Logistic Regression, Random Forests, Support Vector Machines, and variants of Neural Networks. Extensive feature extraction, resulting in a large number of features, can lead to suboptimal results. Learning algorithms under-perform when the number of features is higher than the number of samples, in a phenomenon known as *("the curse of dimensionality")* [52]. As a result, researchers try to reduce the number of features before the training, either with feature selection or dimensionality reduction. For example, in a study that aimed to recognize state

changes in bipolar patients [53], data was reduced using Linear Discriminant Analysis (LDA). Other robust approaches include Principal Component Analysis (PCA). In the next section we will explore some of the most powerful approaches currently used for this type of modelling by introducing deep learning.

1.3.2.2 Introduction to deep learning

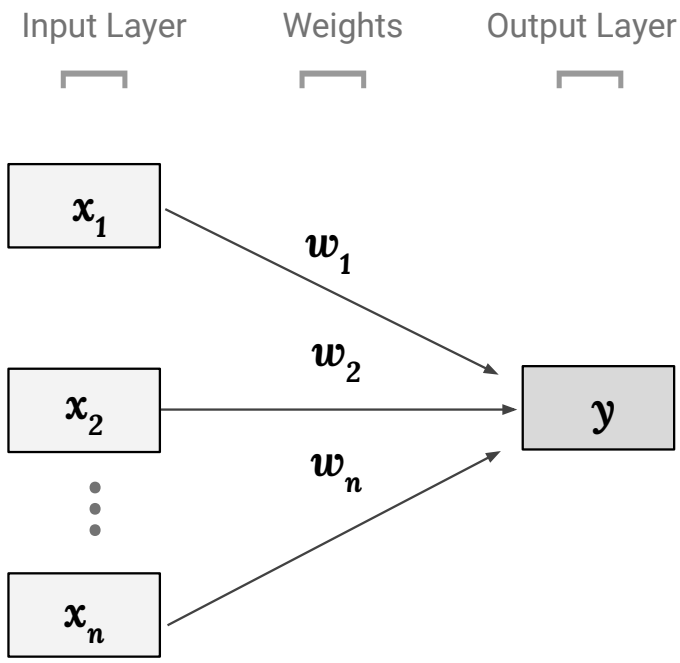
Deep artificial neural networks are the computational engines that power modern AI. To introduce deep learning, we shall start with some of the simple statistical tools: *linear regression* [54]. In linear regression, we take a set of N input-output pairs $\{(x_1, y_1), ..., (x_N, y_N)\}$, for instance Resting HR to VO_{2max} observations. Linear regression assumes that there exists a *linear* function mapping each $x_i \in \mathbb{R}^Q$ to $y_i \in \mathbb{R}^D$ (where y_i may be corrupted by observation noise). As such, regression models the relationships between predictions and targets. The *linear model* in this case results in an linear transformation of the inputs : $f(x) = xW + b$, with W being a Q by D matrix (in $\mathbb{R}$) and b represents a real vector of T elements. In linear regression, different parameters W, b are used to define different linear transformations with the objective of finding the optimal parameter combination to reduce or *minimize* some type of error, for example the average square error over the observed data: $\frac{1}{N}\sum_i \|y_i - (x_iW + b)\|^2$.

Real-life problems are often non-linear in nature, making the use of linear regression to solve them futile. In cases where the relationship between x and y is not linear, we ought to define a *non-linear* function $f(x)$ that maps the inputs to the outputs. To overcome these limitations we can use *linear basis function models*, where input x can be fed through K fixed scalar-value non-linear transformations $\phi_k(x)$ resulting in a *feature vector* $\Phi(x) = [\phi_1(x), ..., \phi_K(x)]$ [55]. Here, the basis function $\phi_k(x)$ is chosen to suitably model the non-linearity in the relationship between the input and the target. As such, it can take many forms parametrized by k such as sigmoidals, sinusoids or polynomials of various degrees. The resulting feature vector ϕ can then be used for regression as opposed to the original data x. A thorough explanation of this process and machine learning from a mathematical perspective can be found in [55].

In 1958 Frank Rosenblatt introduced the *perceptron* model, which is widely considered the first Artificial Neural Network (ANN). Rosenblatt's *perceptron* was a linear classifier that made predictions based on linear prediction functions combining a set of weights with a feature vector. A schematic of the simplest ANN or *perceptron* is provided in Figure 1.3. ANNs or *neural networks* are mathematical models loosely inspired by how biological neurons operate and consist of a set of interconnected model neurons, referred to as *units*. These methods were motivated by the different levels of processing observed in the brain, where it is believed that each level is able to learn different *features* or *representations* at different levels of abstraction [56]. Neuroscientific findings related to the standard model of the visual cortex have inspired the recent developments in the area of machine learning known as *deep learning*. By analogy to real biological neural networks, deep learning models are a form of

representation learning, where the model is fed data and it develops its own representations required for appropriate pattern recognition through multiple layers. Deep neural networks use the ability that *perceptrons* have of representing elementary functions and combine them in a network of layers. These layers are usually arranged in a sequential fashion and composed of a number of non-linear operations, in essence a hierarchical stacking of the aforementioned parametrized basis functions. It is through this process that representations from one layer flow to other layers and are transformed to more abstract representations that allow for highly complex functions to be learned. This ability to adapt with high modularity embodies the true versatility of deep learning models. Indeed, when comparing to their biological counterparts, deep neural networks exhibit some similarities in that they are able to generalize, undergo graceful degradation and are content addressable. An intuitive visual introduction to the concept of manifold learning in neural networks is presented here [8].

Figure 1.3 Example single layered Artificial Neural Network. The connections from the input to the output unit have weights w_1 and w_2.



For the purposes of this introduction to deep learning, we will provide an overview of simple neural network models. Throughout this thesis, more complex architectures are explored, which are introduced in the specific chapters where they are used. In the process of introducing

[8] https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

these simple neural networks we will introduce some of the terminology and notation that will be used throughout this thesis.

Basic feed-forward neural networks:

Rumelhart and colleagues introduced the concept of backpropagation, a process through which the weights of the connections in the network are adjusted so that the difference between the actual output vector of the network and the desired output is minimized [57]. An example network would output:

$$\hat{y} = \sigma(xW_1 + b)W_2$$

given an input x, weights (W_1 and W_2) and element-wise non-linearity σ (which can take the form of a sigmoid, tanh or ReLu to apply the transformation).

This concept is largely based on an application of the chain rule for derivatives. Through the process of backpropagation, the gradient of the loss function $\mathscr{L}$ with respect to the weights w of the network efficiently. For example, for regression tasks a common loss function would be the Euclidean Loss:

$$\mathscr{L}_{Euclidean} = \frac{1}{2N}\sum_{i}^{N} \|y_i - \hat{y}_i)\|^2$$

Where $\{y_1, ..., y_N\}$ are N observed outputs and $\{\hat{y}_1, ..., \hat{y}_N\}$ are the resulting outputs of the model for inputs $\{x_1, ..., x_N\}$.

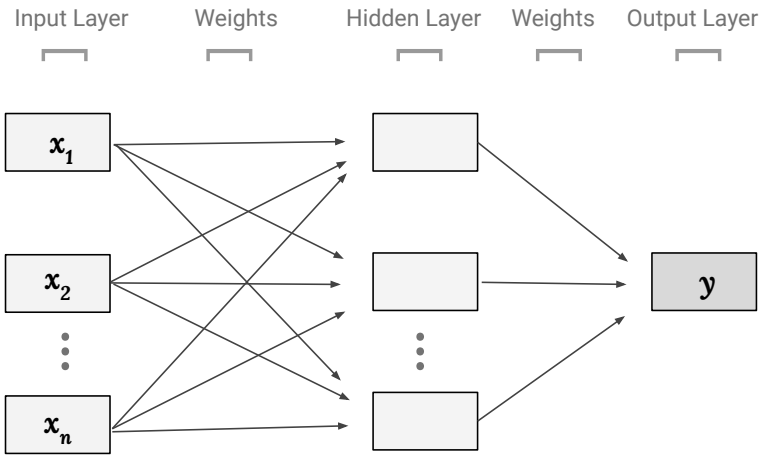
Backpropagation enabled the development of gradient methods, like gradient descent, which allow training multilayer neural networks that are able to distort the input and make it linearly separable through element-wise non-linear transformations (sigmoid, rectified linear units or tanh) that take place in the networks' *hidden layers*. A basic feed-forward model is depicted in Figure 1.4. We will now briefly introduce the building blocks of modern deep learning architectures: convolutional and recurrent neural networks.

Convolutional Neural Networks:

Convolutional neural networks (CNNs) gained popularity due to their strong performance in computer vision tasks, beginning with LeNet [58]. In essence, these models consist of a recursive application of convolution and pooling layers that are then followed by inner product layers at the end of the network. Convolutions are used to extract features from the input data. In the context of images, convolutions conserve the relationship between pixels by learning image features (by applying a transformation) using small square portions of the original image as input data. Pooling layers take the output of such convolution and reduce its dimensionality.

Next, we introduce recurrent neural networks which are optimal for modelling sequential data like text, speech or (as in our case) time-series sensor data.

Figure 1.4 Example Neural Network with a hidden layer.



Recurrent Neural Networks:

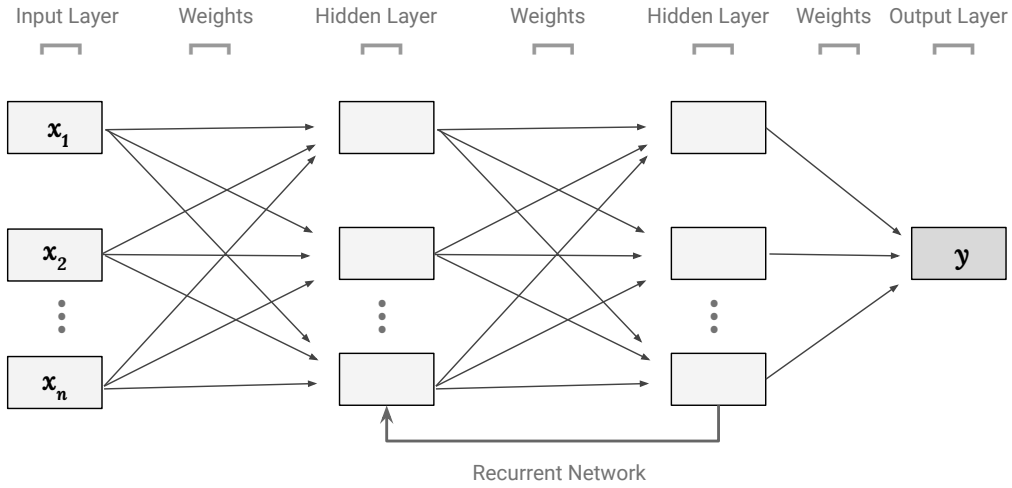
Recurrent neural networks (RNNs) are sequence-based models which allow previous outputs to be used as inputs while having hidden states [59], as depicted in Figure 1.5. RNNs capture historical information of all the past elements of a sequence making them particularly attractive for natural language processing, speech recognition and time-series modelling.

Remark: In our work, various forms of RNNs are used, with Gated Recurrent Units (GRU) and Long-Short Term Memory (LSTM) units being particularly relevant. These models overcome the limitations relating to vanishing gradients present in traditional RNNs [60].

1.3.2.3 Introduction to representation learning

Learning high-level representations from labelled data with layered differential models in an end-to-end fashion is arguably one of the pillars for the success of modern AI. However, a variety of outstanding challenges remain in this field such as data efficiency, generalisation or robustness. Importantly, while supervised representation learning has shown state-of-the-art performance in a variety of real-world applications [61–63], unsupervised learning is yet to see similar breakthroughs as modelling high-level, valuable representations from raw observations (like raw sensor data) remains largely elusive [64].

Figure 1.5 Example Recurrent Neural Network with two hidden layers.



Self-supervised learning

Self-supervised learning is a representation learning paradigm that leverages the intrinsic structure present in the input signals [65]. These models make use of large scale of unlabelled data by learning objectives in order to *get supervision from the data itself*, using supervised loss functions. Through this process, objectives are computed from the signals themselves by applying known transformations to the input data. Most importantly, the intermediate representations capture semantic and structural meaning that can then be exploited for a variety of downstream tasks. While there might be conceptual overlaps with unsupervised learning, its goal is primarily to learn a good geometrical structure of the data using the reconstruction loss (how well we can compress and decompress the input) so that the result can be used mainly for clustering purposes. On the other hand, self-supervised learning's goal is to leverage relationships between the input data using supervised losses (classification or regression). This new paradigm has been successfully applied in filling the blanks for image datasets [66], next word prediction [67], video frame order validation [68], and more recently to small-scale activity data [69]. In BERT [67], for example, the task at hand is to predict the next word given the past sequence. BERT outperforms any other language modelling method by adding two auxiliary tasks within its architecture, both of which are based on self-generated labels. On the other hand, the pre-trained task of video frame order validation [68] improved the performance on the downstream task of action recognition when used as first step. More recently, SimCLR has demonstrated a simplified and improved version of previous approaches to self-supervised learning on images yielding state-of-the-art classification results with a limited amount of class-labelled data in the Imagenet dataset [70].

Throughout this thesis, we use the concept of self-supervised learning to leverage the multi-modal, unlabelled nature of modern wearable sensors.

1.3.3 Human Activity Recognition

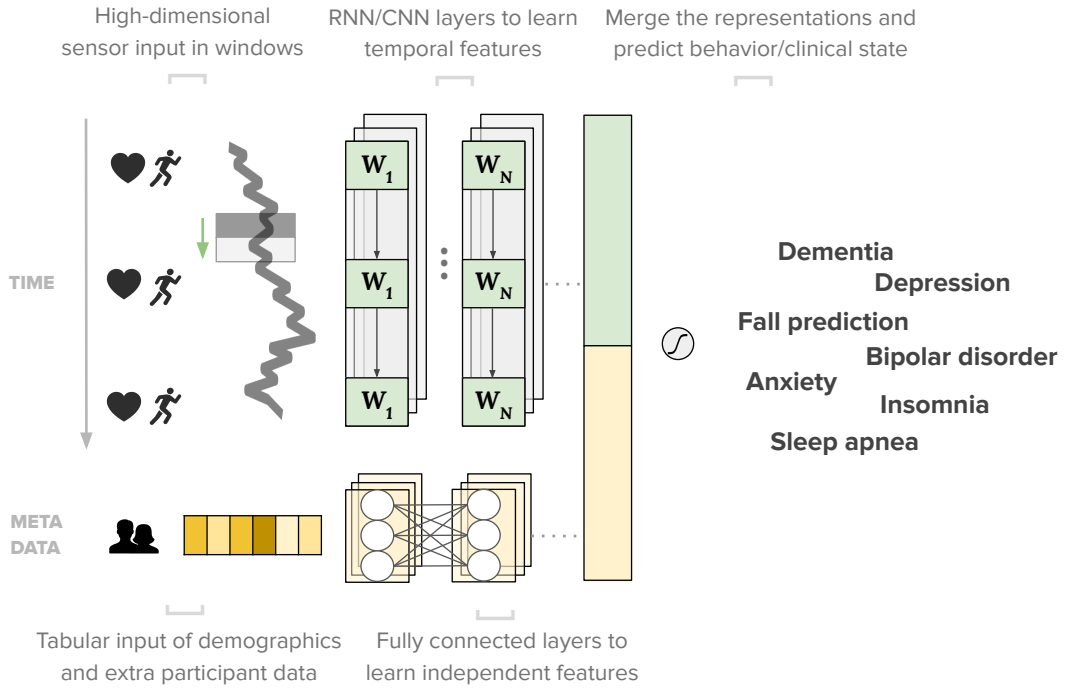
Human Activity Recognition (HAR) can be defined as the challenge of recognizing when a person is engaging in certain activities. Hence, HAR attempts to identify the activity being performed by an individual, alongside when the activity is taking place. HAR systems are based on observations of activities that are captured using a variety of sensors, such as accelerometers and gyroscopes (to capture movement-related data), heart rate monitors (to study heart rate variability) and more. These on-body sensors allow for truly ubiquitous and continuous monitoring of physical behaviors.

The sensors record temporal data, which means that automated HAR methods face a dual issue. First, the method needs to be able to localize contiguous portions of data relevant to the activity recognition problem that the system is facing (*segmentation*). Second, those segments are then classified by automatically assigning class labels. Indeed, this task is particularly complex as information regarding the activity is typically required to identify when the activity took place. However, classification requires previous localization within the sensor dataset to determine when the activity starts and ends. Importantly, the classification step of HAR cannot retrieve any segments that are not included in the original segmentation step, making the task particularly challenging. Due to this dual problem, researchers in HAR often use *sliding-window* approaches to avoid missing any important information for the classification step. The sliding window approach works by providing a small analysis window that shifts along the continuous data stream, extracting contiguous portions of sensor readings. The resulting data is then analysed in isolation, showing strong results in identifying periodic activities such as walking, cycling or climbing stairs. The performance of this analysis largely depends on how the sliding window was defined (length, steps, etc.). Thus, domain knowledge is an important factor when considering the configuration of the sliding window.

Once this process is complete, machine learning pipelines pre-process the sensor data extracted from the sliding window and proceed to extract features and employ probabilistic classification back-ends that are able to assign activity labels to the corresponding analysis window [71]. Over the last decade, deep learning approaches have been established as valuable alternatives to conventional machine learning models which have limited ability to perform in the context of challenging pattern recognition tasks, such as the ones used in HAR. Deep learning models eliminate the need to manually construct feature spaces by automatically learning (*hierarchical*) data representations that are integrated into an overarching classification model. Furthermore, their modelling power has yielded very impressive results as a result of their ability to learn extremely complex decision functions. This is of great importance when dealing with the challenging analytical problems introduced in HAR tasks [72].

Combining multiple sensors for activity recognition purposes has shown promising results. These multimodal approaches have the ability to capture information that may not be possible to explore through individual sensors, such as contextual changes or social interactions [46]. In

Figure 1.6 Multi-modal sensing modelling with deep neural networks. Sensor data is modelled with time-aware layers whilst participant variables are fed into a separate sub-network. The network is trained end-to-end and learns joint representations of both modalities, leveraging latent combinations of sensor features and demographics. (RNN/CNN: Recurrent/Convolutional Neural Networks)



the next section multimodal sensing is introduced, addressing the opportunities and challenges associated to integrating multiple sensors for digital phenotyping.

1.4 Remaining challenges in free-living inferences of physical behaviours addressed through this work

Whilst the mobile and wearable sensing technologies and their associated inferences discussed throughout this introductory chapter continue to progress at a rapid rate, there are several areas where important challenges remain. These provide an important rationale for the work conducted in this thesis.

Indeed, recent technological advancements allow for wearable devices to integrate a plethora of sensors, better batteries and even perform complex in-device computations [73]. However, most devices, their applications and inferences remain poorly validated or not validated at all. Several challenges are worth addressing in the context of wearable device validation. First, to date, commercial device providers rarely provide details of the methods and algorithms used for these inferences, severely limiting the transparency of these products. This impedes the potential use of these technologies for digital health or academic applications as it does

not allow for scrutiny of the methods or results that constitute the backbone of the inferences. Second, commercial devices provide inferences reported to high-levels of granularity, for instance with regards to sleep stages, energy expenditure or fitness. However, academic research has shown the limitations of these claims [74–76]. It is paramount that the inferences that these devices claim are subject to thorough and systematic analytical and clinical validation [77]. Finally, domain adaptation is required to generalise the inferences to other contexts or to slightly different devices. Most wearable inference derivations have been carried in constrained laboratory environments, do not include representative multi-ethnic or cross-cultural samples and are typically constrained to a single device and device placement, severely limiting generalisability to free-living conditions. These shortcomings regarding the diversity of research samples have been previously addressed in other health-related domains such as the treatment effects of specific drugs and how clinical trials should be designed [78, 79]. For instance, it is important to understand how new ubiquitous sensing tools may perform in different ethnic groups, recent work by Bent and colleagues explored the role of skin tone in PPG recordings [80], or more broadly where the representation of minorities in samples matters (i.e.: differences in digital intervention treatment effects).

In this section we will introduce two specific areas where there remains plenty of room for improvement and validation of wearable inferences, these comprise sleep and cardiorespiratory fitness monitoring. Further, we introduce a third area, domain adaptation, which is not covered in the results chapters of this thesis but the author is actively working on and believes is one of the most important challenges for the field of wearable and mobile sensing.

1.4.1 Challenges in the estimation of sleep using wearable sensors

As described, physical activity data is proliferating in both academic and commercial settings with the increasing availability and adoption of wearable devices. However, how to use this data to infer physical behaviours, such as exercise, sedentary behaviours and sleep, remains an ongoing challenge.

As explored in detail in Chapter 2 of this thesis, actigraphy has been used to infer sleep as a non-invasive and free-living alternative to polysomnography for decades. However, two important challenges remain outstanding.

First and foremost, well-established and validated heuristic algorithms for wearable sensors like the Sadeh [81], Cole-Kripke [82], Scripps Clinic [83] or Sazonov [84] approaches were developed for use during the night period. This poses challenges for the evaluation of sleep in free-living conditions at scale when sleep diaries or expert annotations are not present because the original algorithms require a pre-defined search window. Further, the use of sleep diaries in general, as well as for this purpose, has proven to be problematic. These self-report tools are prone to recall bias, with survey data not proving sufficiently reliable labels [85]. When

using sleep diaries, it can often take more than 6 recorded days to achieve agreement with objective labels, even amongst those with more regular sleep patterns [86]. Diary-independent algorithms that do not require intersecting sleep diary data will mean that further improvements in technology do not need to be limited by the feasibility of keeping long-time records and could help answer clinical questions about less regular sleepers such as night-shift workers [87]. Chapter 4 introduces a HR based model for the estimation of sleep windows where sleep classifiers can then be applied.

Second, although commercial wearable devices claim to be able to classify sleep stages, little work has been done to showcase the validity of these devices against gold-standard measures of sleep. Similarly, no study to date has explored the strengths and limitations of leveraging the multimodal nature of current wearable devices for sleep stage classification. This is explored in Chapter 3 of this thesis.

It is noteworthy to mention that, while these new wearable technologies are invaluable tools for the objective monitoring of sleep in free-living conditions, the diagnostic of certain sleep disorders such as insomnia is likely to still depend, at least in part, on self-report measures such as the Epworth Sleepiness Scale (ESS). Thus, we would envision that in the near future digital health solutions for the diagnosis, management and intervention of sleep disorders would likely consist of both objective monitoring of sleep through one or several of the devices presented in Chapter 2 as well as an app or resource that allows the user to input their own perceived sleep health and receive recommendations through this platform.

1.4.2 Challenges in the inference of cardiorespiratory fitness using wearable devices

Cardiorespiratory fitness (CRF) is prospectively associated with the incidence of type 2 diabetes, cardiovascular disease, various types of cancer [88–90], as well as all-cause and cause-specific mortality [91, 92]. Despite its importance, routine clinical measurement of CRF is rare due to its reliance on graded exercise testing, which is costly and unsafe for some patient groups. RHR could serve as a viable alternative to clinical CRF measurement; it is easy to measure, scalable to large populations through wearable sensors, and prospective associations are similar to those reported for CRF [93–98].

In recent work [99], we showed that RHR explains a quarter of the variation in VO_2max among adults in the Fenland study. About half of this association is explained by BMI and PA, suggesting changes in CRF achieved through altered behaviour can be tracked through changes in RHR. This association was verified in longitudinal analyses, making RHR a suitable biomarker of CRF that can be used at a population-level through serial measurement of RHR facilitating personal goal setting/evaluation and remote patient monitoring. The results of this work are discussed in more detail in Chapter 7.

While the inverse association between RHR and CRF is interesting, we hypothesized that the strength of the association could be improved by leveraging the time-series sensor data used in the Fenland cohort. This type of data is similar to that captured by most modern wearable devices. While some commercial devices do provide inferences of VO_2max, most approaches are proprietary and require an exercise recording set by the participant which relies on GPS. Further, these devices often lack proper validation against gold-standard measures of VO_{2max} or derived their inferences in very small samples.

In this thesis we explored how deep representation learning approaches can be used to leverage information from multimodal wearable devices (Chapter 6). First, we showed that forecasting HR and HRV using movement as a pre-training task in a self-supervised learning set-up could effectively be used to generate physiologically meaningful representations. These representation were then used in downstream transfer learning tasks.

For the prediction of CRF, in Chapter 7, we use a deep learning model that learns latent-state representations that emerge from the sequential and dynamic responses of HR and HRV to movement and physical behaviours. These physiologically relevant representations are used to infer VO_2max. We demonstrate that those learned representations not only yield better performance than traditional and state-of-the-art non-exercise models, but also that this paradigm can adapt with time and given behavioural changes by showcasing strong performance in the same population 7 years later. These inferences do not require the participant to record any exercise period manually or rely on GPS and were validated using one of the largest submaximal VO_2max protocols available to date.

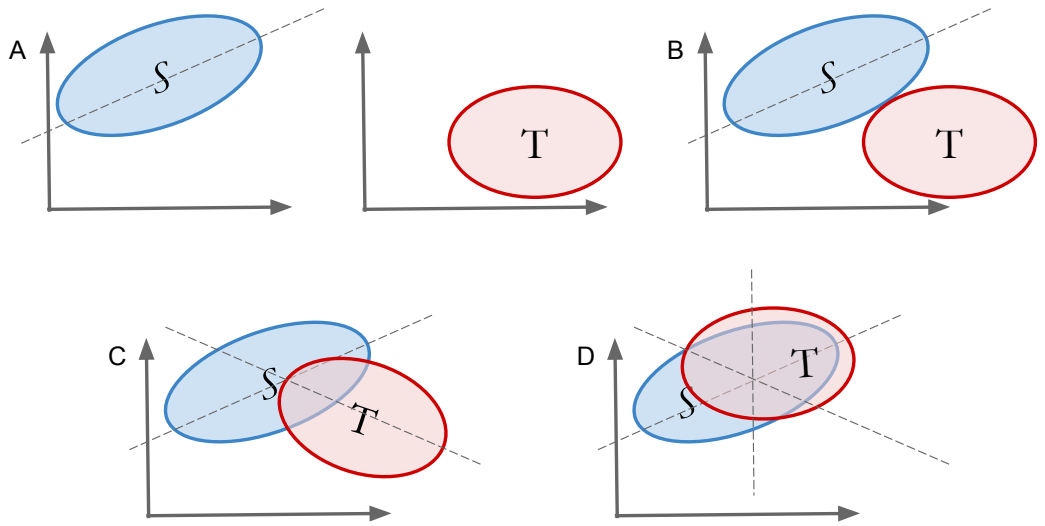
1.4.3 Challenges in Domain Adaptation with wearable sensing

Transfer learning and domain adaptation refer to the paradigm where something is learned in a source domain and is exploited to improve generalizability in another domain.

Differences between source and target data sets for human activity classification and other wearable sensor-related tasks can significantly reduce classification accuracy. Ideally, a model trained on a particular wearable device would be seamlessly applied to other devices that may have a different sampling rate or are slightly different in nature. To achieve this, we turn to *domain adaptation*, a discipline that seeks to develop learning algorithms that can be easily ported from one domain to another, from one device or brand to another [100]. Domain adaptation techniques have shown strong results in natural language processing, as well as image and video classification tasks [101, 100, 102]. More recently, these techniques have been used on wearable and mobile devices in the context of HAR tasks [103–105].

Most existing approaches implement this philosophy of alignment by minimizing a measurement of distributional discrepancy in the feature space, often some form of maximum mean

Figure 1.7 Visual explanation of domain adaptation. Here, we visualize how representations may be aligned in the feature space. In (A) we observe the distributions of a Source (S) and target domain (T). In the S domain an example classifier is applied. (B) Without any correction, we show that the classifier developed in the S domain doesn't generalize well to a context where both the S and T datasets are present. (C) We introduce a potential solution to this issue by training a shared representation to support a source classifier in both domains which can then b used to align the S and T domains across one direction. Finally in (D) we hypothesize that by including multiple self-supervised tasks and learning a number of shared representations, the alignment of the S and T datasets is much more robust and the original classifier derived in S can thus generalize to domain T.



discrepancy (MMD) [106, 107], or a learned discriminator of the source and target as an approximation of the total variation distance [108].

While MMD and learned discriminator approaches are valuable, they rely on the formulation of the training objective as a min-max optimization problem (adversarial learning) which can be complex and computationally costly. One potential solution to overcome the fluctuations observed in the discrepancy loss and divergence attributed to these methods is through the use of self-superevised pre-training tasks in both source and target domains simultaneously. This an area of research in which the author of this thesis is actively involved and a brief explanation is featured in Figure 1.7.

1.5 Thesis rationale and aims

The overarching aim of this PhD is to advance understanding of how multimodal sensing through wearable and mobile devices can be used to characterize physical behaviours in free-living conditions. In this introduction, we explored the state-of-the-art in mobile and wearable

sensing, providing an overview of the foundations to some of the fundamental methods and techniques used in this field. We also established the importance of accurate characterization of physical behaviours, including sleep, sedentary behaviours and exercise, to understanding their relationships with health and disease, and ultimately driving the personalized and public health recommendations of the future. Overall, to date, the modelling methods described throughout this chapter have helped the field move towards the collection of large-scale data from free-living conditions with unprecedented granularity. This is likely to have important implications for industrial and academic purposes, facilitating the types of well-powered epidemiological studies of physical behaviours, discussed in Section 1.2.

Despite this progress, some important outstanding gaps in the existing literature remain that motivate the studies reported in this thesis. These will be addressed through four interconnected sub-aims:

- **Aim 1: To estimate sleep and sleep stages in free-living conditions through multi-modal wearable sensing.** The methods used to characterize physical behaviours require further refinement to ensure their accuracy. This particularly pertains to the study of sleep. As noted, whilst multimodal sensing has the potential to elucidate free-living physical behaviour on a hitherto unprecedented scale many longitudinal studies still rely on sleep diaries [109, 110] which have been shown to be subject to recall biases [85]. Further, in cases where objective sensor data is used, existing algorithms designed to infer sleep have rarely been validated against gold-standard measures [32]. Existing algorithms typically pertain only to one type of device limiting the extent to which they can be used with ease across studies. Additionally, to fully leverage multi-modal sensing, the merging of information from different sensor modalities is needed, as single modality approaches may have limited inferential capabilities for specific behaviours. An example of this is that while actigraphy signals have been used to infer sleep-wake periods, they are not capable of distinguishing sleep stages when used in isolation [111].
- **Aim 2: To explore how wearable sensors can be used to better understand and characterize physical activity beyond intensity metrics.** Conventional wearable devices have focused on reporting counts (actigraphy) and intensity metrics such as VM or ENMO (accelerometry) [16]. However, the widespread adoption of triaxial accelerometry, including in cohorts such as the UK Biobank [112], today enables a more nuanced analysis of physical behaviours. For example, the analysis of postural changes made possible through wearable sensing makes it possible to better understand sedentary behaviours that traditional, intensity based approaches may have missed [113]. It is helpful to infer detailed information about physical activity, because characteristics of different activities, beyond intensity, may have important consequences for health [114]. In order to explore these relationships, we first need to be able to accurately and comprehensively characterise activity.

- **Aim 3: To leverage multimodal free-living sensing information to infer meaningful physiological characteristics, including fitness or BMI.** Novel uses of multimodal wearable sensor data need to be explored to make optimal use of existing information. For example, wearable sensors may be able to be used to infer health-related characteristics, such as CRF. As these characteristics have known associations with health and disease [115], being able to infer them in situations when they are not directly measured will enhance our ability to further interrogate their health relevance and adjust for them in studies of other traits. A limited number of published works have begun to do this, for example [116, 46]. However, thus far, they have primarily focused on laboratory-based activity recognition tasks and have not explored health-related inferences in free-living conditions in detail. In other fields, these methods have shown great promise. For example, in computer vision multimodal representation learning has lead to valuable improvements in audio-visual speech classification [117].
- **Aim 4: To summarise and provide recommendations for the future management of mobile sensing data and digital phenotyping.** The number of individuals collecting their own data through personal devices and contributing data to research has increased substantially, for instance, the use of wearable devices is projected to increase more than five-fold over five years from 2016[9]. Unlike for other specialised types of personal data, such as genetic information [118], there is no governance framework for digital phenotyping information. Indeed, most digital phenotyping data arises from commercial products, where the role of this data and the associated research is, at least in part, designed to support a business model. Most of this data is therefore not used to produce pure public goods or knowledge and is not freely available under existing governance frameworks for proprietary data. The current fragmented approach to regulatory oversight, classification of data for the purpose of identifying the applicable laws and varying data governance practices together decrease user trust in digital phenotyping and limits potential medical research. Thus, there is a clear and growing need for a governance framework that protects the interests and rights of both individuals and researchers.

The aims of this thesis are addressed, as outlined in Figure 1.8 through Chapters 2-8.

In order to answer these outstanding questions and address the aims of the thesis, we turned to some of the largest objective monitoring cohorts available to date with gold-standard measures of sleep, energy expenditure or cardiorespiratory fitness that enable the validation of our methods. These include: the Fenland Study, the Multi-Ethnic Study of Atherosclerosis and the Biobank Validation Study among several others. A full description of these cohorts, as they are relevant to the thesis, is given within each results chapter.

[9] Wearable Devices https://www.statista.com/study/15607/wearable-technology-statista-dossier/ (Statista, 2019)

Figure 1.8 **Elaboration of the aims of the thesis.**

CHAPTER 2

UBIQUITOUS MONITORING IN SLEEP HEALTH: A DATA-DRIVEN REVOLUTION IN SLEEP SCIENCE AND MEDICINE

Publications

Parts of this chapter are published elsewhere:

Perez-Pozuelo, I., Zhai, B., Palotti, J., Mall, R., Aupetit, M., Garcia-Gomez, J. M., ... & Fernandez-Luque, L. (2020). The future of sleep health: a data-driven revolution in sleep science and medicine. NPJ digital medicine, 3(1), 1-15.

Contributions

I planned this project and recruited a working group of global experts on sleep health. I wrote this chapter and the resulting manuscripts.

2.1 Summary

In recent years there has been a significant expansion in the development and use of multi-modal sensors and technologies to monitor physical activity, sleep and circadian rhythms. These developments make accurate sleep monitoring at scale a possibility for the first time. Vast amounts of multi-sensor data are being generated with potential applications ranging from large-scale epidemiological research linking sleep patterns to disease, to wellness applications including the sleep coaching of individuals with chronic conditions. However, in order to realise the full potential of these technologies for individuals, medicine and research, several significant challenges must be overcome. There are important outstanding questions regarding performance evaluation as well as data storage, curation, processing, integration, modelling and interpretation. In this chapter, we introduce the state-of-the-art in sleep monitoring technologies and discuss the opportunities and challenges from data acquisition to the eventual application of insights in clinical and consumer settings. Further, we explore the strengths and limitations of current and emerging sensing methods with a particular focus on novel data-driven technologies, such as Artificial Intelligence which is readily used in the results chapters of this thesis.

2.2 Introduction

Sleep is a crucial biological process and has long been recognised as an essential determinant of human health and performance. Whilst not all of sleep's functions are fully understood, it is known to restore energy, promote healing, interact with the immune system and impact upon both brain function and behaviour [119, 120]. Even transient changes in sleep patterns, such as acute sleep deprivation, can impair judgement and cognitive performance, whilst long-term aberrations have been linked to disease development [121, 122]. Global trends in sleep suggest a decrease on average sleep duration [123–126]. Given these trends and the implications of sleep for health and well-being, better characterisation of sleep characteristics represents a public health priority [127–129].

Chapter Significance: In this chapter, an overview of the state-of-the-art and upcoming sleep sensing technologies is presented, highlighting the strengths and limitations of each on of these tools. This chapter also introduces the digital sleep framework and outlines the major challenges associated to this field. Importantly, wearable devices, which are used throughout this thesis and in the next following chapters for sleep sensing purposes, are of particular interest given their low user burden and relatively good accuracy.

Sleep is known to be regulated by three main factors: circadian rhythms, sleep-wake homeostasis and cognitive-behavioural influences [119]. With regards to behavioural determinants, poor sleep quality [130] (as defined by the National Sleep Foundation's recommendations based on total sleep time, sleep latency, wake after sleep onset, number of awakenings $>$ 5 minutes and sleep efficiency) has been associated with stress, anxiety, smoking, sugary drink consumption, workplace pressures, financial concerns, regularity of working hours, physical activity, sleep regularity and commuting times [127, 131]. Indeed longitudinal research has linked changes in physical activity to changes in the severity of sleep-disordered breathing and, hence, disturbed sleep [132]. Further, dietary patterns have shown associations with sleep quality [133]. It is now understood that the associations between diet, physical activity and sleep are bidirectional. Thus, poor sleep, high levels of inactivity and a poor diet comprise inter-related public health priorities [134]. The mental and physical impairments associated with a single night of poor sleep can outweigh those caused by an equivalent lack of exercise or food [135].

Sleep loss affects every major system in the human body. Chronic changes in sleep have been associated with a plethora of serious medical problems from obesity and diabetes to neuropsychiatric disorders [131, 136, 137]. For example, chronic insomnia is associated with both incident cardiovascular disease and all-cause mortality [138, 122]. A 2011 meta-analysis of prospective studies, which included 470,000 individuals, explored the association between sleep duration and cardiovascular disease [139]. Relative to those who slept between 7 and 8 hours per night, those who slept less than 6 hours exhibited a 48% increase in the incidence

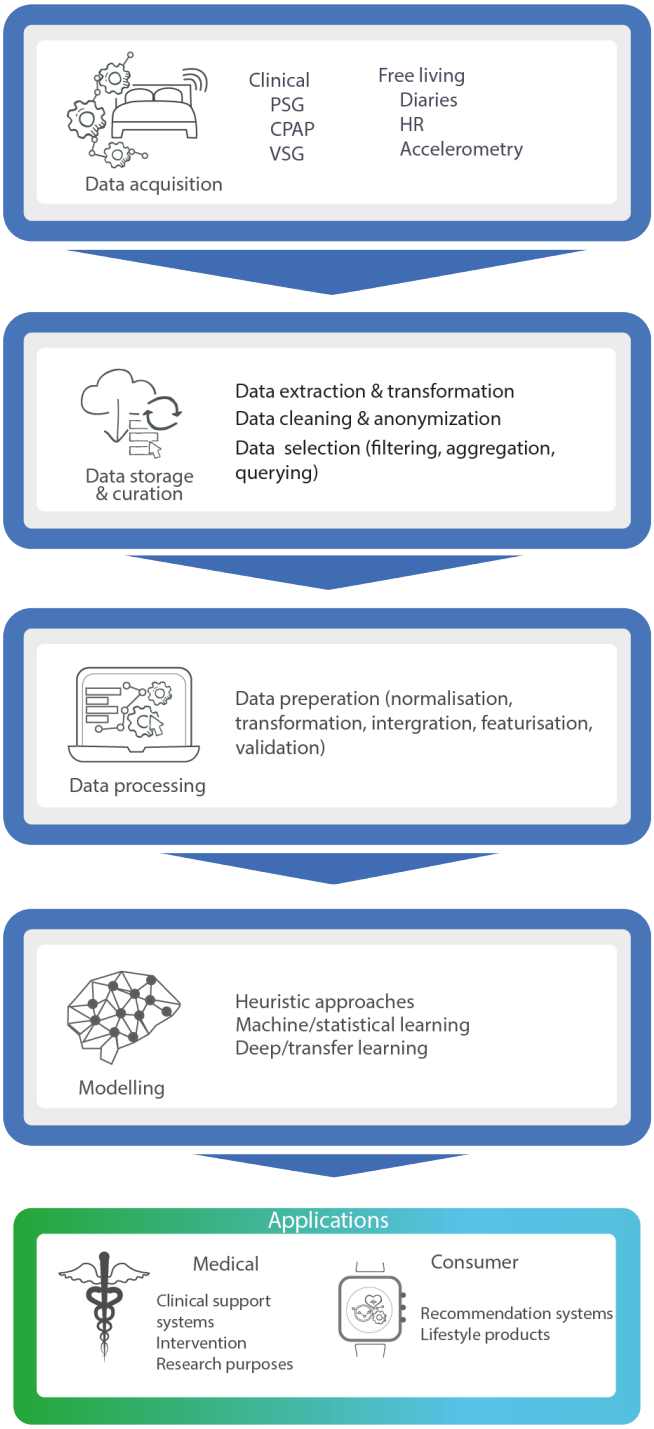
of coronary heart disease and 15% increase in the incidence of stroke, whilst those who slept greater than 8 to 9 hours exhibited a 38% increase in coronary heart disease, a 65% increase in stroke and a 45% overall increase in cardiovascular disease [139]. Other large epidemiological studies have also reported associations between sleep and cardio-metabolic disease, including reports studying the effects of shift-work [140–143]. Short sleep duration has further been associated with incident diabetes and weight gain, as well as impaired appetite control [144, 136]. Shortened sleep and poor sleep quality have also been identified as risk factors for cognitive decline, neurodegenerative disease, mood changes and depression, as well as other neuropsychiatric conditions [145–148]. There is also mounting evidence linking sleep to both immune function [149] and cancer [150, 151]. In a seminal study published in 2002, Spiegel and colleagues demonstrated an association between sleep deprivation and a muted immune response to flu vaccination [152].

Besides its ramifications for the health of individuals, sleep has macro-level economic implications. A recent study estimated the annual economic cost of poor sleep to the Australian population at $45.2 billion, comprising direct health care costs, the cost of associated health conditions, reduced productivity, accidents and informal care [128]. Moreover, in a 2016 report, RAND Corp quantified that the combined cost of insufficient sleep across 5 OECD countries (Canada, USA, UK, Germany and Japan) exceeds $600 billion a year [127].

Following mounting evidence of the role of sleep in well-being, its relationship with disease and mortality and its economic impact, there has been increased interest in measuring sleep characteristics. This has led to an expansion in the development and use of sleep-related technology. In particular, recent developments in digital technologies designed to improve the measurement and characterisation of sleep have demonstrated particular potential. These advances facilitate the objective and unobtrusive measurement of sleep characteristics in large, free-living populations at scale, facilitating well-powered epidemiological investigations designed to explore the relationships between sleep and disease [32]. Furthermore, these developments are set to have clinical implications for the monitoring and diagnosis of sleep disorders and, ultimately, could be used for the modulation of sleep [153].

The core objective of this chapter is to discuss the implications of digital technologies for the study, monitoring and modulation of sleep. To aid this discussion, we introduce a 5-step *Digital Sleep Framework*, which comprises the complete process from sleep data acquisition to end-user applications of insights. Figure 2.1 depicts the framework. This chapter is structured around the framework's 5 steps: data acquisition; data storage and curation; data processing; modelling; and applications. Finally, we discuss the biggest challenges and opportunities in this field followed by conclusions based on our findings.

Figure 2.1 The digital sleep framework covers the path of sleep data from its acquisition to when its insights are used for medical or consumer applications. The framework begins with the acquisition of sleep-related data. This can be done using a variety of sensors. This data is then stored and curated, a step that comprises privacy-aware storage, cleaning, filtering and anonymization. Once that data has been appropriately treated, the processing step takes place whereby data is transformed and integrated based on the end-model. For example it may undergo different transformations like normalization or featurization. The next step entails modelling, which can consist of simple heuristic methods, statistical learning or deep learning methods, for example. Finally, the resulting model can be deployed for a variety of either medical or consumer applications.



2.3 Sleep data acquisition

Sensors have been used to study sleep for decades. Traditionally, polysomnography (PSG), paired with clinical evaluation, has been the gold-standard and de-facto technique to study sleep in clinical and laboratory settings as well as to diagnose a subset of sleep disorders [154]. However, in recent years, industry and academia have invested heavily in the development of smaller, less obstructive and more portable devices for the continuous monitoring of sleep [155]. This is motivated by a desire to enable data acquisition in larger participant groups over more extended periods and in a more natural setting by decreasing both the cost of monitoring and the burden to participants. However, challenges remain in data acquisition, including the provision of ubiquitous, less obtrusive and stigmatising long-term acquisition mechanisms. Moreover, long-term patient monitoring usually suffers from missing periods that may mislead the estimations of health markers. Here we discuss the current state-of-the-art of sleep data acquisition in clinical and free-living settings, including an overview of traditional and novel approaches and their strengths and weaknesses.

Traditional sleep monitoring and monitoring in laboratory settings

Since the 1960s, **polysomnography** (PSG) has been used in clinical settings to monitor sleep through a battery of simultaneous, complementary sensors [156]. These sensors typically allow for the measurement of: (1) brain activity through electroencephalogram (EEG), (2) airflow, (3) breathing effort and rate, (4) blood oxygen levels, (5) body position, (6) eye movement, (7) electrical activity of muscles and (8) heart rate. Traditionally, PSG requires participants to sleep in a laboratory setting. The results are then scored by an expert who has received training on how to interpret these signals. Ambulatory PSG is an alternative modality which often uses a reduced number of sensors and allows monitoring to occur at home, outside of the laboratory. This facilitates the monitoring of patients with disorders that may not be easy to evaluate in a laboratory setting [157].

To-date PSG remains the gold-standard for sleep measurement. However, the technology is limited in its use as it remains impractical for long-term home use. This precludes its use in long-term sleep monitoring or sleep in free-living settings beyond the laboratory. Furthermore, PSG is expensive, time consuming and requires trained technicians to administer and interpret. As a result, the scalability of this technique for large-scale population-based studies is very limited, particularly when the aim is to assess typical sleep patterns in free-living, naturalistic conditions. Whilst Ambulatory PSG provides a partial solution to some of the issues, it remains both expensive and burdensome.

Another conventional method used in clinical settings to evaluate sleep is **Videosomnography** (VSG). VSG encompasses a range of video-based methods used to record a person as they sleep. These video recordings are subsequently used score sleep behaviours. VSG has been typically paired with PSG in clinical settings to study sleep disorders. However, recent advances in

telemedicine have made the use of home VSG increasingly possible. Although VSG is typically scored by experts in a time-consuming manner, advances in signal processing and AI have led to the new possibility of automatically-scored VSG [158]. However, VSG presents similar scalability issues as PSG. It is costly and, at present, requires expert monitoring and scoring.

Data on sleep can be also obtained from devices to treat sleep apnoea such as **Continuous Positive Air Pressure** (CPAP). These devices can be used both in laboratory and home settings. For example, Aggarwal and colleagues showed that CPAP can be used to classify and track sleep metrics which could be used to monitor the response of CPAP therapy in sleep apnea patients [159].

Sleep monitoring outside the laboratory.

To understand the role of sleep in health and disease sleep must be monitored in a free-living environment and in a non-obtrusive way to ensure the sleep captured is as representative of typical sleep as possible. As such, low-cost, wearable sleep detection systems are a promising tool to study sleep architectures in free-living individuals at a population level. At present, there are several options for the monitoring of sleep outside the laboratory. These comprise actigraphy, heart rate sensing and other wearable technologies. Multiple published works have demonstrated that a single modality sensor representation, such as heart rate alone, is not sufficient to accurately complete sophisticated sleep stage classification [160]. The availability and range of digital technologies for the measurement of sleep has significantly expanded in the last decade. Both consumer and medical grade devices across a variety of fields (wearable, remote sensing, mobile health, clinical-grade) have become more sophisticated and affordable. Nevertheless, comparing performance across different platforms and methods remains a challenge and few methods have been validated against gold-standard PSG or undergone systematic reliability assessment [161].

Traditional free-living sleep sensing and measurement approaches: actigraphy and accelerometry.

Actigraphy and accelerometry are non-invasive methods to monitor human activity and rest cycles. They have been used to describe physical activity levels in large scale populations [162] and can also be used to monitor sleep. These methods offer an affordable, scalable alternative to PSG to monitor sleep-wake cycles and have now been recognised by the American Academy of Sleep Medicine as a valid method for the assessment of sleep [163]. Recent advances in AI and larger studies in conjunction with PSG have resulted in the refinement of the method [164]. However, three key limitations of actigraphy and accelerometry remain. These are: (1) the lack of validation studies for the different consumer-grade devices (2) lack of standardisation of approaches for human-activity recognition, and (3) the lack of assessment techniques for daytime sleeping. Nowadays, wearable sensors can be used in combination with other minimally invasive sensors (such as heart rate monitors, miniaturised ECG, pulse-oximetry, blood pressure monitors, galvanic skin conduction, light sensors, gyroscopes, barometric

Figure 2.2 Emerging Sleep Sensing Technologies. Emerging sleep technologies range from non-contact methods like RF sensors to miniaturized, wireless or in-ear EEGs.

altimeters and GPS trackers). Sleep can also be monitored through a combination of wrist actigraphy, hip sensors, smart phone sensors and under-mattress sensors [165]. Nevertheless, this increased availability of sensors also results in a greater challenges when optimising the match between the end-application and the sensor used [161]. A description of actigraphy specific sleep metrics is provided in the supplementary material.

Emerging sleep sensing technologies.

The fundamental aim of ubiquitous computing in sleep tracking is to achieve miniaturisation of sensors and non-intrusive sensing that can pervasively monitor physiological signals related to sleep activities. Embedding different types of ambient sensors into objects that we interact daily is more attractive than using multiple redundant sensors collecting homogeneous information. Embedded devices, such as bed sensors, have been developed to track different sleep-related metrics such as sleep time, breathing, snoring, heart rate, body and room temperature or humidity levels [166–168]. Whilst these sensors are interesting and potentially valuable for clinical and epidemiological research, as well as wellness and sleep education, very little is known about how their performance against gold-standard measures and more research is required to evaluate their usability. Some have emerged in recent years but remain at an early stage of development (e.g., WiFi and radio-signal approaches), whilst others have been around for longer (e.g., smartwatches). They are depicted in Figure 2.2 and discussed below. Some of the potential techniques to unobtrusively measure sleep through the acquisition of physiological signals include the following:

Bed sensors: Bed sensors may be defined as any sensor that sits on the bed and can be used for monitoring physiological processes. Body movements, breathing and even cardiac activities can be detected by the volume change of the pneumatic underneath an individual whilst they are lying in bed [169–171]. For instance, using micro-bend fibre optic sensors underneath the mattress allows for monitoring of breathing and body movement activities that can be then used to extrapolate some valuable sleep metrics [172]. Similarly, fibre-optic based systems have allowed not only for the analysis of different motion types but also for the introduction of retroactive feedback based on those movements [173]. Unobtrusive sleep monitoring using bed sensors (either on the mattress or the bed frame) usually entails monitoring of movement, but also respiration rate and occasionally heart rate. Several companies, including Apple (Beddit), Nokia and Withings, have released new sensor accessories that can be attached to a person's mattress or bedframe and often interact with a separate mobile application or dashboard. Nevertheless, a range of determinants can influence the performance of these methods, from postural differences to inter-subject variability in BMI and pre-existing clinical conditions [171].

Consumer-graded wireless EEG and reduced-array EEG: EEG is an integral part of PSG and is also used in a variety of neuropsychiatric tests and applications. Conventional EEG requires expert set-up and can be burdensome, uncomfortable and is not portable. Wireless EEGs have gained traction in recent years, with several established companies, as well as start-ups, launching products. Their performance for sleep monitoring has been compared to conventional EEG that is part of PSG and has demonstrated strong results [174, 175]. Further, Koley and colleagues showed that automatic-scoring using ensemble models on a single channel EEG could yield agreement rates of 0.87 when compared to expert scoring of the same signal [176]. Whilst this study was conducted in a clinical environment, and hence lacks the recording conditions required for free-living validation, together these investigations show that the results of conventional EEG can be approximated by simpler devices that may be able to be scored automatically.

Similarly, several miniaturised EEG devices have shown promising results with regards to their ability to classify sleep stages [177, 178]. In-ear EEG is a modality that has shown promise in recent years, for instance, Mikkelsen and colleagues compared in-ear mobile EEG analysed through machine-learning-based automated scoring to conventional, manual scored PSG and commercial-grade actigraphy showed promising results, although also constrained to a laboratory environment [179]. A 2019 study showed that automatic sleep stage prediction based on a single in-ear sensor demonstrated a 74% agreement with the hypnogram generated from full PSG, which is promising but still requires further work for it to be at a clinical standard of performance ($\approx$ 90% agreement) [180]. These devices are particularly interesting given their potential for free-living application. They also have the advantage of conserving much of the granularity and information that a conventional PSG-based EEG would offer in non-laboratory set-ups [181].

Although the performance of these wireless, miniaturised and in-ear EEG devices is promising, more extensive studies are required to determine the feasibility for use in population science and in a free-living environment as well as for applied sleep research studies.

Smartwatches and fitness trackers: A plethora of wearable smartwatches and activity bands have been developed to infer sleep. These devices often derive their metrics using a combination of movement signals (accelerometry, as explored in previous sections) and heart rate and heart rate variability. Henriksen and colleagues assessed the validation or reliability of some of the most common brands on the measurement of physical activity and sleep (Fitbit, Garmin, Misfit, Apple, Polar, Samsung, Withings and Mio) [75].

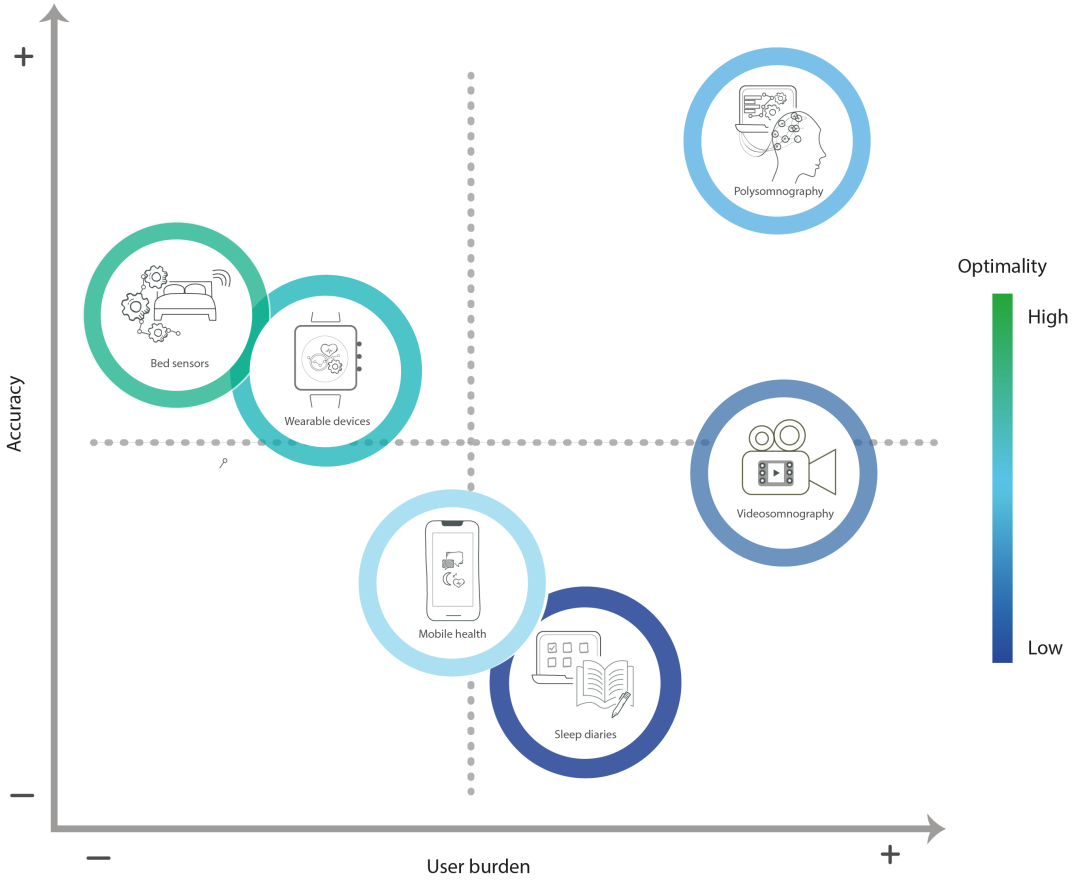
Mobile phone sensing: Mobile phones offer a wide range of sensors, such as gyroscopes, microphones and accelerometers, that can be used to monitor sleep patterns [182]. For instance, iSleep, developed by Hao and colleagues, leverages a smartphone's built-in microphone to detect events that happened during sleep such as body movement, cough, and snoring by processing the acoustic signals [183]. The software achieves accuracy of over 90% for event classification (snoring, cough, sleep) under different environmental conditions. An important limitation of the system is that the high-rate microphone sampling represents a significant source of energy (and battery) consumption.

Several other sleep applications can be found on the different app stores these days. Sleep Cycle is among the most popular ones, using both accelerometry and the built-in microphone to track sleep and provide personalised alarm clocks, waking up the users at ideal timings (during light sleep) [184].

Ultrasound sensors: Ultrasound sensors can be used to detect body movement and breathing patterns during sleep [171, 185, 186]. These sensors provide information regarding the frequency and type of body movement through the Doppler technique. This technique mirrors that used in conventional radar systems and allows the retrieval of parameters related to breathing rate, heart rate and body motion. The method has been shown to detect physical movements with an 86% recall rate and error rates of less than 10% [187]. The most pressing limitations of this method are, however, the fine-tuning required based on the type of targeted body and the sensitivity to small movements [188].

WiFi and radio signal approaches: In the past decade, high frequency and sub-millimetre wavelength radio technologies have demonstrated the ability to capture physiological signals without body contact. The principle is to send a low-energy radio wave towards an individual who is in bed and then to detect the signal bounced back from the body. Through signal processing, it is possible to extract biological information such as breathing patterns, heart rate and full-body motion from these findings [189–191, 187]. These biological signals can be used to determine sleep stages as shown by Zhao and colleagues [192], as well as to monitor insomnia [193]. The main challenge with this approach is that the signal is subject to a lot of 'noise' and the information related to sleep needs to be extracted. Moreover, the measurement

Figure 2.3 Selected methods for the measurement of sleep and their accuracy and usability trade-off. This chart plots the accuracy of sleep sensing methods at inferring sleep-related metrics against their ease of use. For example, while polysomnogry is considered the "gold-standard" technique to measure sleep, it is cumbersome and expensive.

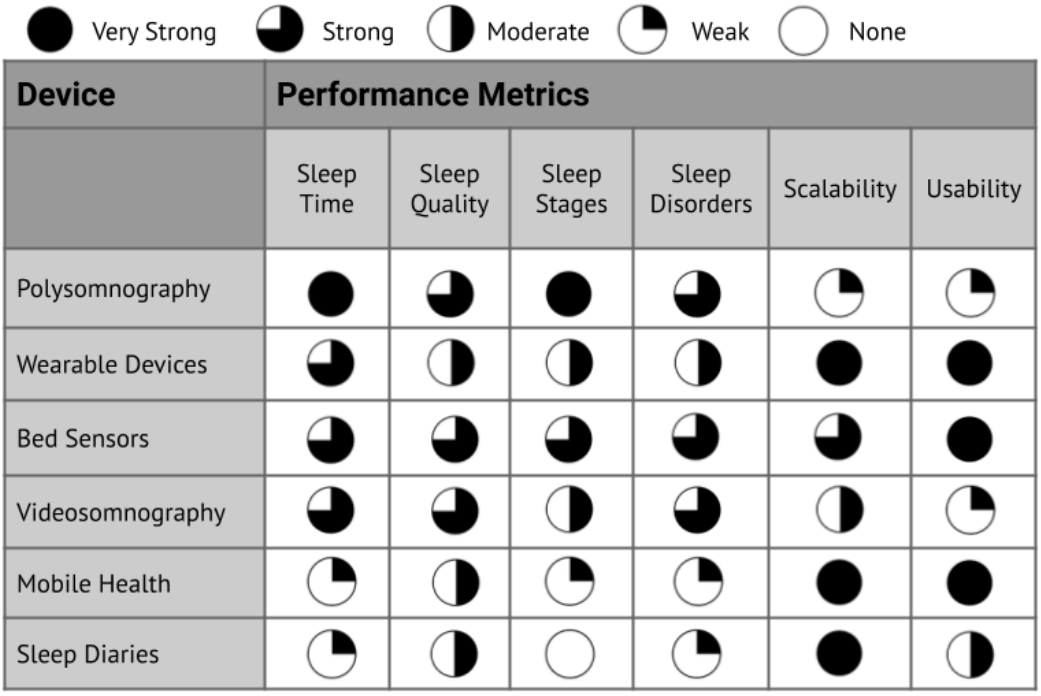


conditions are also strongly dependent on the individuals being monitored. In particular, the signal reflects all objects in the bedroom and is affected by the sleeping position of the individual [194].

Some of the methods described in this section are, in general, more accurate or more usable than others. Figure 2.3 shows a scheme of the accuracy versus usability trade-off for the main methods described in this section.

Data collected from different modalities representing diverse physiological information may have varying predictive power and noise topology as explored in Figure 2.4. However, different modalities and the information they collect may be highly complementary and, in practice, aggregating sleep data from various sources may make models more robust and tolerant both to noise and missing data. Such complementary fusion protocols have been shown to significantly improve the classification performance of sleep stages [195, 196].

Figure 2.4 Holistic evaluation of sleep-monitoring methods. Some methods, such as PSG, are accurate but inappropriate for use in daily sleep monitoring, as they require professional set up and are intrusive. Other methods, such as bed sensors, are unobstrusive but more prone to noise than PSG.



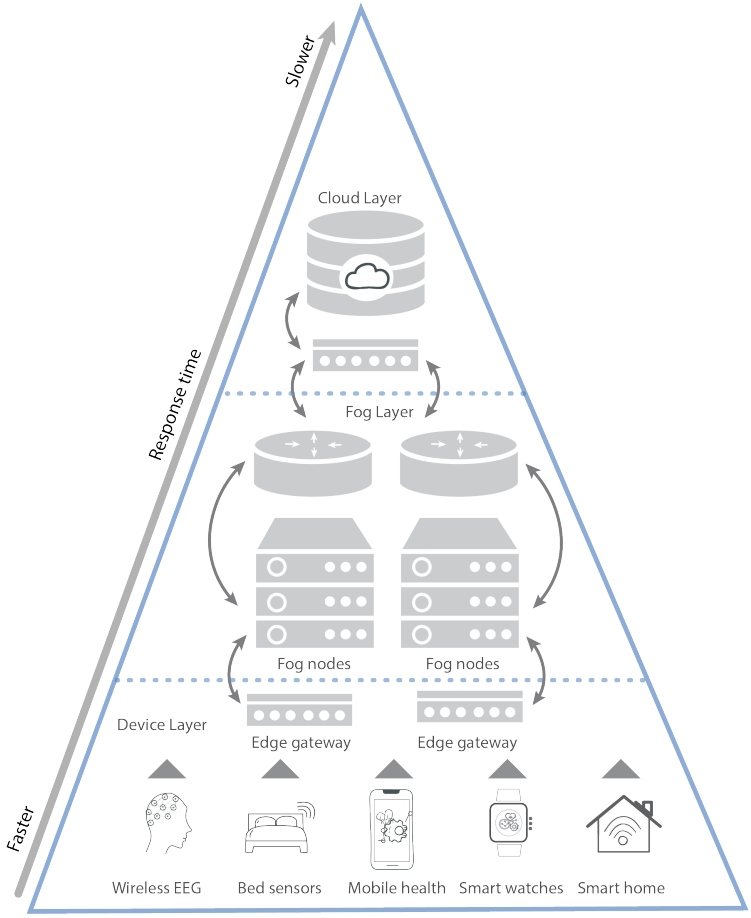
2.4 Sleep data storage and curation

Regardless of its intended end-use, all the data collected using the methods and sensors previously discussed requires appropriate storage, curation and processing prior to analysis. Until the turn of the century, analogue PSG systems, limited to analogue amplifiers and paper tracings, were common practice for storing sleep information. However, with the development of digital recording systems, these types of analogue recordings have become outdated as different challenges have emerged for handling data from new digital sleep technologies. For example, in the era of digital medicine, systems often require real-time storage and processing of data collected as part the so-called Internet of Things (IoT) [197] and Big Data Analytics [198]. IoT links all sorts of connected devices into comprehensive networks of inter-correlated computing intelligence without with need for human input. With regards to sleep, the integration of IoT technology has several challenges. These include data storage, management and exchange across different devices and sensors, alongside privacy, security and data access concerns.

Cloud computing integration with IoT is gaining traction in healthcare and is being used for digital sleep applications. For instance, 3-layered architectures composed of (1) an IoT layer sensor acquisition/data compilation; (2) a Fog computing layer for event processing; and (3) a Cloud layer for data management and Big Data Analytics have been proposed for sleep

monitoring use cases that integrate several sensors [199]. In Figure 2.5, an overview of data acquisition and the movement of information from sensors to the cloud is explored. The remainder of this section discusses the fog computing and cloud storage layer more fully.

Figure 2.5 Overview of cloud-computing based sleep data acquisition and storage. This illustration provides an overview of the process starting with device layer (which includes fast, real-time processing and data visualisation, embedded systems, gateways and micro data storage), followed by the fog layer (which includes local networks, virtualisation, data analysis and reduction) and finally cloud layer (which consists of data centres and big data storage and processing)



Fog computing layer

Fog computing entails data analysis on edge devices, which enables real-time data processing, reducing costs and also improving data privacy. Fog computing is commonly deemed mini-cloud computing as it performs all the processing locally. The fog computing layer abstracts the heterogeneity of the incoming data formats, communication technologies and protocols from the sleep-sensor IoT layer. Platforms, such as Smart IoT Gateway, have emerged as solutions to communicate with all the heterogeneous IoT sensors potentially deployed in home environments and perform local processing before transmitting the data to the cloud layer [200]. Fog computing seeks to achieve a seamless continuum of computing services connecting the cloud to the devices (IoT). This contrasts to edge computing which isolates and keeps

the computing at the "network edges" [201] and facilitates the aggregation of multi-modal physiological data from different devices and sensors that are then processed locally (e.g. processing data directly on an IoT Gateway). This architecture can provide near real-time decision-making to support sleep monitoring and intervention.

Following the receipt of signals from the devices, pre-processing at the fog computing layer includes 3 main operations: (1) the fusion of signals provided by different IoT sensors; (2) detection of periods containing missed data; and (3) imputation of missed data. When sleep sensor signals contain missing data, it is usually because the user did not wear or was not in contact with the sensors. However, functional errors can also occur. For example, smartwatches may run out of battery or memory and may fail to communicate with the user's smartphone. Missing data can be detected by various algorithms, including through simply thresholding the smoothed signal.

Besides data pre-processing, the fog computing layer also enables the inter-operability of heterogeneous sources of the data. Inter-operability is a key function of Smart IoT Gateways. It allows for communication and integration of devices which are operated on different protocols and use different technologies. Furthermore, the Gateways facilitate the sharing of information and the driving of actuators or components that meet the required needs of the system [202]. For instance, it can be used to detect sleep apnea events and activate motors designed to change the users' body position or to play sounds or music during particular sleep stages [203].

Cloud storage layer

Cloud computing architectures include servers, networking, software, databases and data analysis over the internet which enable fast deployment, flexibility and economies of scale. Cloud computing is often considered the centralised paradigm, while the fog computing layer previously described would be a decentralised paradigm. Nevertheless, as explained in Figure 2.5, they can effectively work together.

Sensor data integrity is paramount for successful application and analysis in digital medicine and requires appropriate data storage in order to be realised [204]. Relational databases can be limited when storing and analysing semi-structured data obtained from multi-modal sleep sensing technologies. Hence, current trends to store and query digital sleep data are based on Not Only SQL (NoSQL) databases such as MongoDB, Cassandra, HBase or CouchDB, which allow for better representation of heterogeneous data structures and batch data. Moreover, several of these NoSQL databases provide connectors to cluster-computing frameworks, such as Apache Spark, Storm, Flink and Hadoop, enabling Big Data Analytics. These Apache products are a good fit for both batch processing and stream processing via in-memory computation and processing optimisation [205]. Resilient Distributed Dataset (RDD) allows Apache Spark to simultaneously store data on memory and write to storage media based on pre-defined criteria from the real-time data stream. Hadoop allows for batch processing and the use of MapReduce algorithms to analyse data stored in Hadoop Distributed File System (HDFS). HDFS can

handle petabyte level data analysis, which can be used to provide in-depth statistical analysis of clinical sleep data and can also be used in large population epidemiology studies.

2.5 Data pre-processing

Before sleep data can be used for modelling, it must be pre-processed. As discussed in the preceding sections, there is a growing trend toward the integration of sleep data from various sensors. As such there is a preponderance of unstructured multi-modal time-series data with substantial noise. For example, different equipment brands and models may be equipped with different quality of sensors, amplifiers and electrodes that result in different noise topology as a result of their unique materials and manufacturing process. Data measurement, processing and storage may also differ between sensors. For example, depending on the application and device, it might store RR interval instead of raw ECG. Hence, data needs to be cleaned and filtered, removing artefacts that differ depending on the modality employed before any feature extraction or modelling can take place.

Depending on the nature of the data, several pre-processing approaches can be applied. Smoothing and de-noising can remove unwanted spikes, trends and outliers from a signal [206]. For example, polynomial de-trending methods can remove continuous quadratic or linear trends that may be caused by impedance changes on the skin. Similarly, Hampel filtering can remove unwanted spikes from sinusoidal signals. Noise arising from other sources should also be be considered. This may include power line interference, thermal based resistive changes or contact conductive artefacts. These noises can be filtered by applying various band-pass filters. The ultimate objective of de-noising is to ensure that the noise is subject to a specific distribution, such as a Gaussian distribution, as far as possible.

Beyond de-noising and smoothing, re-sampling and standardising can be used to improve data integrity and consistency in the pre-processing stages. Linear or higher-order interpolation can be used to fill missing or corrupted data, as well as for data scaling, through methods such as linear scale-transformation [207]. These methods can suppress the noise levels and variability in the signal and transform the data into a pre-defined range without altering its distribution. Data standardisation, such as min-max standardisation and z-score standardisation, can suppress noise levels and variability in the signal and transform the signal such that it approximates a normal distribution.

2.6 Artificial intelligence-based sleep modelling

Once sleep data has been pre-processed, data modelling can be commenced for different applications. Today, many of these modelling and application tasks are based on AI, which

entails the use of algorithms and techniques that mimic human cognitive functions, reasoning and problem-solving skills and have brought a paradigm shift to digital medicine. Indeed, the influence of AI in medicine is growing rapidly and is being exploited in a variety of fields from clinical medicine to population studies [208]. In essence, the application of AI in medicine aims to aid clinical decision-making through analysing complex medical data. The insights generated can then be used in diagnosis, treatment, the prediction of clinical scenarios and to aid scientific discovery [209]. Increasingly, AI is changing research methodology and facilitating the personalisation of medicine through its advancements [208].

With regards to sleep science, the impact of AI is multifaceted. First, it can aid clinicians in making sleep disorder diagnoses [210]. This is achieved by translating collected sensor data into pre-defined knowledge (e.g., class label), providing an inexpensive and objective alternative to manual sleep stage scoring [211]. Similarly, through its automated analysis capabilities, AI can provide wellness and lifestyle recommendations based on the interpretation of data collected from wearable devices and mobile apps [212, 213], enable clinicians and researchers to track changes in sleep patterns from people's homes [214] or interact with smart-home set-ups to provide better quality sleep through the adjustment of lights and temperature in rooms [215]. Here we discuss methods of AI-based sleep modelling.

AI applied to sleep science.

Traditional AI systems were rule-based, requiring the programming of pre-conceived rule sets and demonstrating limited flexibility. By contrast, machine learning (ML) provides a more flexible alternative to data modelling, especially when applied to the raw unstructured signals. In plain terms, ML aims to train, learn and optimise a mathematical model which can transform or map the collected (complex) signals into comprehensible knowledge.

Usually, ML approaches, which include logistic regression, support vector machines and random forest, tend to use structured data as input. This makes feature engineering or feature extraction a standard procedure before model training. Feature engineering can be achieved in various forms. For example, given a sliding window (from the raw time-series data), statistical features such as mean, standard deviation, energy, entropy and so on, or time-frequency features such as wavelet/Fourier transform coefficients can be extracted and used as input for the traditional ML models. Moreover, in some applications, domain experts can also design features based on their understanding of the signal in certain fields. Compared with the raw signals, the engineered features tend to be low-dimensional with information redundancy suppressed, making the model training tasks more effective.

From a ML perspective, the most common tasks for sleep research are the classification of sleep-wake cycles and stages as well as the derivation of sleep-wake metrics. Although heuristic approaches and some traditional ML approaches have demonstrated reasonable performance in some tasks, the feature engineering process tends to be time-consuming, and may require domain knowledge in some circumstances, making the whole system design an expensive

process. On the other hand, the new methodologies offer more flexibility in sleep modelling. For example, deep learning methods can be used to perform end-to-end training, which directly maps the raw signal into the target labels. It is a pure data-driven process, and latent patterns can be automatically learned without the feature engineering process.

In Table 1 , we highlight several popular ML models that can be applied to different sensing modalities.

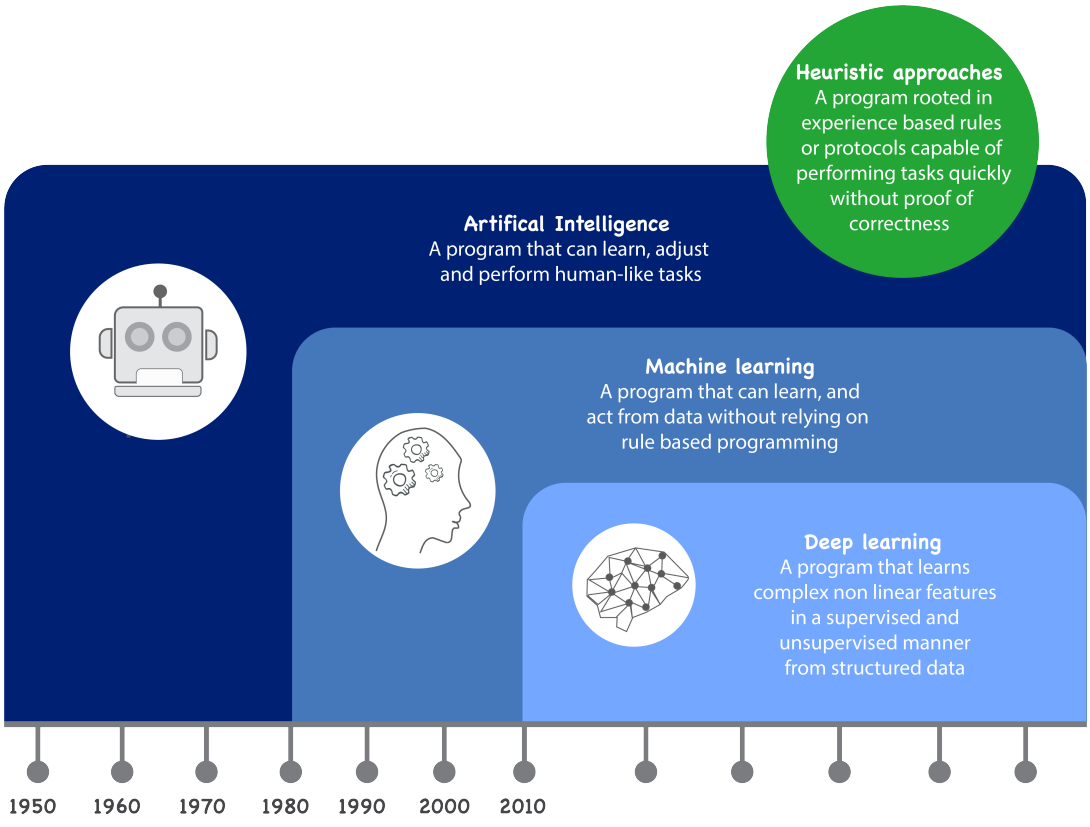
Conventional sleep classification methods.

Following the American Academy of Sleep Medicine (AASM) guidelines, traditional sleep scoring in neurophysiology laboratories assigns 1 of 6 labels to each 30 second epoch. These are as follows: (1) awake; (2) rapid eye movement sleep (REM); (3) non-rapid eye movement (Non-REM); (4) sleep stage 1 (N1); (5) sleep stage 2 (N2); and (6) sleep stage 3 (N3). This task is performed manually by trained sleep technicians based upon data generated through PSG. Sleep stages themselves are associated with physiological changes that are useful for the diagnosis and assessment of specific sleep disorders such narcolepsy [216]. For example, respiratory monitoring in PSG facilitates the detection of sleep-disordered breathing, such as obstructive sleep apnea. In this disorder, abnormal breathing events are less severe in N3 than N1 sleep due to the change in central control of breathing, and more severe again during REM given upper airway muscle tone reduction [217].

Manual sleep scoring suffers from several draw-backs. It is time-consuming, subject to biases, inconsistent, expensive and must be done offline. Rosenberg and colleagues reported that the average inter-scorer reliability for sleep stage scoring was approximately 83% [218]. This estimate is similar to that reported in other studies [219]. By contrast, the use of AI and automated sleep stage classification algorithms represents a fast, non-subjective, inexpensive and scalable alternative to this traditional sleep scoring approach. Aside from issues of reliability, it can take 1-2 hours for an expert to score a night of clinical PSG recordings [220] while the automated system can finish the same task in seconds. Thus, multiple approaches and methods have been used to distinguish sleep from wake automatically as well as to characterise specific sleep stages. In broad terms, sleep classification algorithms fall into different categories, but these categories can be closely interrelated, as shown in Figure 2.6. The different categories comprise traditional algorithms and both ML and deep learning approaches. These are elaborated below.

Traditional algorithms for scoring of sleep from either PSG or actigraphy signals tend to be based on heuristic approaches [221]. These heuristic approaches are themselves based on prior knowledge of the sensing modality and sleep physiology. In actigraphy, the use of the magnitude feature as a proxy of movement for sleep/wake classification provides one example. This approach offers quick solutions with fast implementation but also tends to be biased by the programmer's understanding and interpretation of the problem and to perform differently depending on the population in which they are applied. For example, the algorithm

Figure 2.6 Sleep classification algorithms can be based on heuristic approaches or Artificial Intelligence. We describe machine learning/statistical learning approaches and deep learning approaches within AI.



developed for a nocturnal sleep pattern may not suitable for non-nocturnal sleep. Penzel and colleagues provided an in depth review of some of these approaches in clinical settings, offering a quantitative analysis of their performance and requirements [222]. Palotti and colleagues evaluated the performance of some of the most common approaches, including statistical ML, on actigraphy data [223].

Machine learning and deep learning approaches have gained traction in recent years for the task of classifying sleep-wake cycles and sleep stages in multi-modal sensor data [224, 225]. With the availability of raw actigraphy signals, several deep learning techniques such as convolutional neural networks [226] and recurrent neural networks [227] have been used to exploit the temporal nature of this unstructured data to distinguish the sleep-wake cycles [223] robustly and understand the role of activity in sleep related disorders [228]. While the evaluation of most traditional and ML algorithms are performed using standard quality metrics such as accuracy, precision and recall per class, it is also important to measure clinically relevant metrics such as waking after sleep onset (WASO) and sleep efficiency [223]. By optimising clinical metrics, ML methods enable the physicians to make informed, clinically relevant decisions. Adequate performance defined by quality metrics varies depending on the task intended. For instance some sleep disorders may not require high-levels of granularity for

their diagnosis whereas interventions that aim to boost *deep sleep* ought to rely on accurate granular classifications of sleep stages.

Table 1 2.1 provides a holistic overview of the most common classification methods based on the modality used (PSG/EEG, wearable device (accelerometry/actigraphy), Others (heart rate/PPG/etc)). References are provided for methods by modality in the appropriate cells. It is important to note that different methods are ought to be used based on the objective at hand. For instance, deep learning methods often provide better performance than traditional statistical learning methods, but require large computational power and lack the interpretability that other models offer [223]. Performance and model evaluation is discussed in further detail on the supplementary material.

Table 2.1 **Sleep classification techniques across different sleep sensing modalities**

Technique	Technique Variations	PSG/EEG	Wearable Sensing
Statistical	Latent Dirichlet Allocation	[229]	
	Support Vector Machines	[230]	[231]
	Hidden Markov Model	[232]	[233]
	Quadratic	[234]	
	Bayesian	[235]	
	Logistic Regression	[236]	[237]
Instance Base	K-Nearest Neighbors	[224]	[225]
Decision Tree	Decision Tree	[238]	[32, 239]
Ensemble Model	Adaboost	[240]	
	Bagging	[241]	
	Random Forest	[242]	[243, 244]
	XGBoost	[245]	
Clustering	K-Means Classifier	[246]	
	Spectral Clustering GMM	[247]	
ANN and DNN	Convolutional NN	[248]	[237, 192]
	Recurrent NN (LSTMs, GRUs)	[249]	[237, 250, 251]
Others/Heuristic	Fuzzy Classifier	[252]	
	Wavelet Methods	[253]	
	Sadeh		[163]
	Sazonov		[84]
	Oakley		[254]
	Cole-Kripke		[255]
	Webster		[256]
	ADAS		[257]
	Scripps Clinic		[83]

Emerging approaches for sleep classification.

There are a plethora of methods available for the predictive modelling of sleep-related problems, as mentioned in previous sections. However, several outstanding questions remain regarding their application. Issues such as model sustainability, handling the heterogeneity of data and variability in the demographics, behaviour and lifestyle of the population and generalisation to unseen data, need to be investigated more comprehensively. Below we highlight some of the emerging technological solutions for the handling of these issues.

Model sustainability. An important consideration is that the majority of existing ML models perform a task (such as sleep-wake classification and sleep-related disorder prediction) by learning from an underlying distribution of data. However, in real-world conditions, the data generated from participants can change over time due to age, lifestyle changes, new sensing modalities, the progression of sleep/health disorders or other changes. An imperative question then is how to make the trained model sustainable in response to changing domains. Life-Long learning might be the first step to address some of these challenges. This would facilitate sustainability by allowing the model to evolve over time [258].

Personalised sleep classification. One of the major challenges that AI encounters when facing sleep classification tasks is inter-subject differences. That is, the intra-class variability (e.g., differences in length of REM sleep between participants) can be too large to be captured by the trained model, making inference process prone to errors. By contrast, by taking personal information into account, a human analyst can address this problem easily. In the case of PSG scoring, a skilled neurophysiologist may consider demographic characteristics (such as age and gender) and adjust their scores accordingly. Despite advances in AI methods for sleep classification across different sensing modalities, most of the current models do not adapt to individual characteristics. There may exist large inter-subject variation and the trained model (on the population-level data) may not be the optimized one for certain individuals.

In the future, personalisation could be a useful approach to improving the performance of the AI-based sleep modelling systems, improving the performance of algorithms [259, 260]. It has been suggested that personalisation may be especially useful when using data from noisy modalities, such as wearable devices [261]. Model personalising has been successfully applied in other fields such as mood recognition [262] and seizure detection [263]. However, it remains relatively untapped in sleep science. Recent works have shown that transfer learning could be used to realise personalisation. For example, based on EEG modality deep neural networks were trained on a large population, followed by a fine-tuning process at the subject-level [261]. The results suggested that substantial performance gain can be achieved [261].

The process of personalisation can also be applied in the aforementioned distributed networking environment. Federated learning proposes a distributed way of updating a centralised model by aggregating each patient's local updates into a central server [264]. The distributed update framework not only provides a model parameter update mechanism but also creates a personalised predictive model by feeding individual data to a global model in the localised updating process.

Generalised sleep classification. Another way of improving the performance of AI-based systems is to reduce the effect of the contextual information for better generalisation [265]. Based on adversarial training process, some of the most recent works performed subject-invariant learning, which makes the system less sensitive to personal and environmental factors [228, 266]. Similarly, by undergoing an adversarial training procedure, temporal dependencies

can be learned. These then transfer well to new subjects and different environments in sleep classification tasks [192]. Pillay and collegues used EEG data in combination with a generative modelling process to obtain agreement between the labels estimated and clinician's labels for automatic 4 stage sleep classification in infants [183]. In general, by learning the representative features that are less sensitive to contextual factors and thus robust in various (complex) sleep classification tasks (such as diagnosing sleep apnea or insomnia), these approaches aim to increase the generalisation capabilities for better performance. This is an alternative to the aforementioned personalisation approaches.

2.7 Data-driven sleep applications

There is a wide and growing range of commercial, health and clinical situations for which data-driven sleep applications are being used. In hospitals, sleep medicine units have traditionally used PSG and more recently actigraphy/accelerometry, for the diagnosis and monitoring of sleep disorders [267]. One of the main challenges in sleep medicine is the increasing incidence of sleep disorders, which in turn leads to higher demand on sleep labs to provide diagnoses. Consequently, software for sleep medicine is being gradually upgraded to include automated sleep-metric calculations and seamless integration of sleep data sources, such as sleep questionnaires. These upgrades have the potential to compliment Electronic Health Records, enabling health care practitioners to better manage their patients sleep disorders [268]. Figure 2.7 gives a brief overview of key areas that will be affected by the impact of sleep technologies and newly generated sleep data.

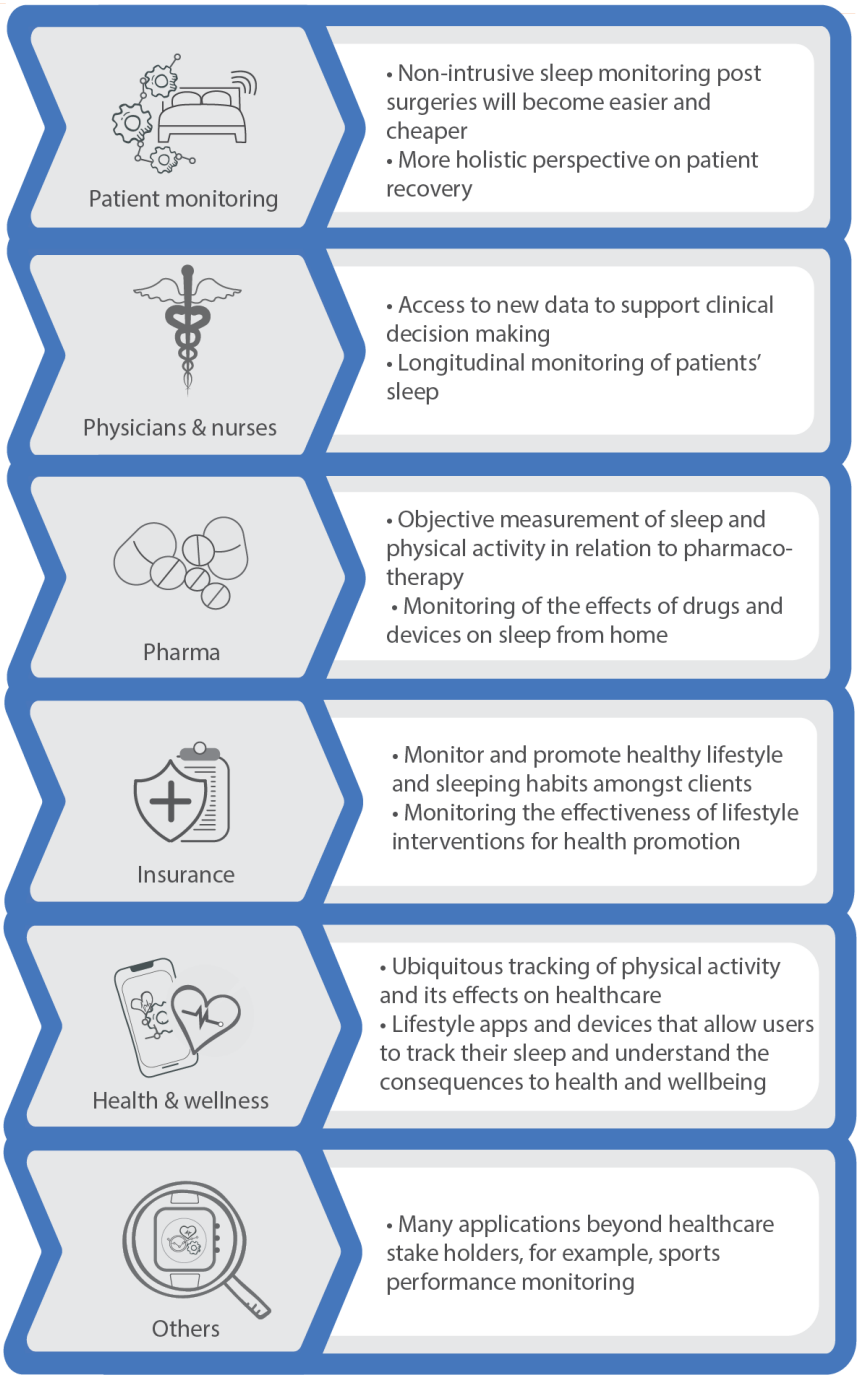
Sleep data in health and disease As discussed, disturbed sleep has been linked to reductions in quality of life and to a higher risk of a plethora of chronic conditions [269, 270]. Thus, it is of vital importance for digital self-management and monitoring solutions to include tools that allow accurate monitoring and assessment of sleep quality. Aside from its direct role in ill-health, poor sleep quality can worsen the symptoms of many serious and chronic conditions including cancer and multiple sclerosis [271–273]. Moreover, pharmacological treatments may in turn worsen sleep disorders as side effects. For example, smoking-cessation drugs and some treatments for cancer have been shown to reduce patients' sleep quality [274, 271]. Due to the complexity of the relationship between sleep and health, there is a need for the design of digital intervention methods to address the unique requirements of sleep within long-term or chronic conditions. There are early examples of mHealth interventions to improve sleep quality on people with cancer and diabetes, amongst others [275–277].

Amongst otherwise healthy individuals, there is also an increasing interest in mobile and wearable applications for health and wellness [213]. Ultimately, it has been proposed that such technologies could be used to direct personalised sleep health recommendations to individual users [153]. Furthermore, other consumer sleep technologies have gained traction in recent years, and although they still need appropriate clinical evaluation, they could enhance patient-

clinician interaction and sleep self-management [278]. Some of the most common commercial and familiar technologies, such as Fitbit or SleepAsAndroid, offer monitoring and tracking sleep quality. In addition, new sensing technologies, such as those discussed in the sleep data acquisition section, are gaining traction and devices like Beddit have attracted investments from large technology companies [279]. These monitoring technologies are complemented by applications aimed at improving quality of sleep by supporting a more suited wake timing using approaches such as smart lighting or smart alarms that only ring when the user is in light sleep. Despite their growing popularity, at present, most of these consumer-oriented technologies lack validation and their underlying models change frequently [213, 278]. This fast growing industry needs to be matched by multidisciplinary scientific efforts that evaluate the performance, usability and value proposition of new sleep technologies.

When exploring the impact of the digitisation of sleep on wellness and health promotion, it is also important to mention occupational health applications. Often sleep disorders are the result of lifestyle factors including, for example, prolonged screen-time before bed. The resulting poor quality of sleep can feedback to that lifestyle, by reducing productivity. Consequently, corporate and health insurance wellness programs are starting to offer incentives and personalised coaching to clients and employees, with some initiatives directly promoting sleep quality at the workplace [280]. For instance, FirstBeat provides a solution for companies that comprises personalised sleep and physical activity monitoring for employees combined with personal face-to-face coaching with the aim of increasing employee health and employment satisfaction [281]. However, these technologies can also be exploited, as in West Virginia prior to the teacher's strike, where declining to wear a fitness tracker and meet a certain step count resulted in a $500 penalty annually for their healthcare payments.

Figure 2.7 Key areas of impact for sleep health. Emerging sleep health technologies will have an impact on patient monitoring, clinical care, insurance, the pharmaceutical industry and health and wellness applications, as well as other impacts including on digital therapeutics and sports performance.



Data visualisation and visual analytics.

Data visualisation, in general terms, is the graphic representation of data. Abstract data is processed such that they can be represented using visual objects (e.g., points, lines, bars, etc.) ease of interpretation and better understanding. Visualisation relies on human's high

throughput visual perception channel, and the ability to connect data representations to human knowledge and expertise which are not encoded directly in the data [282].

Visualising health-related data goes back to the days of paper charts and maps. Since the rise of internet and mobile application, digital displays are ubiquitous and people are now widely educated to read standard graphics representing data. Typically, activity data is presented based on the time component, which is usually visualised using line charts, where the horizontal x-axis is time. Raw signal visualisation is mainly meaningful for domain experts trained and experienced to interpret complex patterns. Specific patterns can be automatically detected or highlighted on the chart, for instance, when activity levels go above or below a threshold [283]. Projection techniques are also a popular means of reducing the dimension of high-dimensional data for better visualisation and knowledge generation [284, 285].

Sleep data visualisation is only meaningful if the resulting visualised data make sense to the end-user, which can be challenging for non-expert users of wearable technology. SleepExplorer is an example of visualisation research aimed at understanding how users can benefit from visualising their own personal sleep data [286]. SleepExplorer organises a flux of sleep data into sleep structure, guides sleep-tracking activities and highlights connections between sleep and other related factors such as napping, coffee and alcohol intake, as well as mood. Recent studies have analysed behavioural change resulting from techniques implemented in activity trackers and their visualisation, but few studies are focused on sleep [287]. Ravichandran and colleagues' conducted a study of user's experience and understanding of sleep metrics provided by sleep sensing devices. Their findings suggest that visual feedback may be helpful to users [288].

However, several challenges remain in regards to sleep data visualisation. (1) Scalability: For large-scale historical health data, visualisation requires adapting to large time scales (from minute-level to year-level information) and displaying meaningful data summaries to the user or primary care practitioner. (2) Heterogeneity: The data collected from different devices varies greatly from, for example, GPS location or glucose levels to pictures of food or phone-screen time. This poses a challenge for the visualisation of personal data for patients and for the health care professional [289, 290]. (3) Usability: Sleep data visualisation should be tailored to end-users and their specific needs [291].

2.8 Challenges and opportunities

With advances in technology, the volume of physiological and clinical data resources available to biomedical research is expanding [292]. This includes open-source data from Electronic Health Records, medical image repositories, genomic archives and massive person-generated data from wearable technologies [292–294]. Recently, sleep data repositories, such as Sleep-data.org, have been created to advance the field [295]. These repositories include multi-modal

sleep data (from clinical-grade PSG to actigraphy and questionnaires) [295] and are being used to create ML benchmarks [223]. These developments are crucial for the creation of generalised ML models that can be applied reliably to clinical and commercial settings to further our understanding of the role of sleep in well-being and disease.

Sleep-related technologies are not only useful for monitoring but may also be used to aid intervention. For example, the portability and pervasive use of mobile phones makes them an attractive option for the delivery of interventions and several studies have already shown promising results when using mobile phone platforms for sleep interventions. These interventions include, but are not limited to sleep advice for behavioural change [296, 297], optimised alarms based on sleep stage [298] and sleep tracking and feedback [212, 299]. Furthermore, new sleep-technologies may be able to complement or augment current clinical-grade diagnostic tools for sleep disorders. A 2017 review by Shin provides an in depth overview of this area of research as well as the strengths and limitations of the current efforts [300].

Despite the potential of technologies and open resources, challenges must be overcome if their potential is to be realised. Their heterogeneity, variability (both between sources and over time) and data quality is, at present, a strong barrier to efficient data reuse. Appropriate analysis also remains a challenge. To overcome this, temporal and source variability of signal repositories must be characterised [301, 302] and common representation spaces should be defined to exploit shared latent information among data distributions. Indeed, appropriately representing data and metadata originating from different sensors (type, make, version, etc) is critical in order to later harmonize and integrate data from disparate sources as well as for sensor data fusion. Models should adapt their inferences from different data sources and at different points in time.

In addition to data handling and analysis challenges, new sensing technologies require systematic validation [303]. These validation requirements vary based on the end-use of the technology and must be held to higher standards if they are to be used in clinical settings [303]. On a population level, there is a wide and growing interest from the general public in wellness mobile and wearable applications, which in many cases are related to sleep and inform people's lifestyle decisions and understanding of their health [288]. Nowadays, there are hundreds of sleep applications and a plethora of wearable devices that claim to track sleep quality [304]. However, most of those devices have little or no information regarding their reliability and validity, the testing they underwent or how the data is acquired (i.e. sampling rates, pre-processing, etc) and processed [305, 306]. As such, individuals could become concerned or reassured about their sleep based on unreliable data. Further, concerns have been raised about the performance of these devices in populations with chronic conditions and mobility problems [307]. Whilst this has, to-date, mainly been confined to concern regarding the tracking steps and physical activity, these devices must be tested in a range of populations, in particular, those with sleep problems. Massive usage of consumer-grade sleep tools may

also increase individual's health concerns and have a ripple effect on overstretched health-care systems [278].

Lack of reliability and validity testing also poses several obstacles to the use of data-driven applications in sleep medicine and research. As explained in a 2016 editorial by Wilbanks and Topol [308], the lack of transparency in these technologies limits researchers' capabilities to study any potential bias due to the lack of information on the characteristics of the cohorts. Further, this lack of transparency makes it more difficult for researchers and clinicians to use these devices and mobile applications. Despite some initiatives, such as Apple HealthKit or C3-PRO [309], which aim to facilitate data sharing across platforms, these data tend to be highly summarised and in a post-processed state. Summary level data is not always appropriate for use in academic research or some AI applications as the processing steps are often not described.

To generate maximum benefit to the end-user or other stakeholders (e.g., hospitals, researchers, public health officials, regulators, industry) there is an increasing need for a safe and effective clinical biomarker ecosystem with algorithmic transparency, inter-operable components and sensors and open interfaces that allow for high integrity measurement systems [310]. This will allow for the verification and validation of digital biomarkers for sleep health.

Finally, there are data-privacy concerns. Sleep tracking mobile applications and wearable technologies often collect information such as movement, GPS location and sound, which could have potential applications beyond the tracking of sleep. These privacy concerns may be mitigated through the deployment of data processing functions on the user's mobile equipment, without requiring server processing [311]. Similarly, an alternative is to empower users to decide what data they want to send to the server [311].

2.9 Conclusions

The impact that sleep has on human health is undeniable. Recent advances in sensing technology, big data analytics and AI allow for truly ubiquitous and unobtrusive monitoring of sleep and circadian rhythms. However, challenges remain to realisation of the benefits of this monitoring for individuals, research and clinicans. Here, we introduced the Digital Sleep Framework, a framework outlining the steps required from the multi-modal acquisition of sleep-related data through to its clinical and commercial application and exploring all aspects of this chain. As the number and scope of sleep monitoring technologies continues to grow and the diversity of digital sleep solutions and applications continues to multiply, the need for careful, risk-based product validation has become increasingly important. The heterogeneity of sensors used for the monitoring of sleep-wake cycles and circadian rhythms poses a unique set of challenges for modelling and interpretability. Hence, the identification and standardisation of robust, reproducible digital sleep biomarkers is of paramount importance. Modelling based

on these signals must be as free as possible from conscious and unconscious bias and the development of algorithms must be transparent and readily available for all stakeholders.

The digitisation of sleep is likely to have repercussions across industry, healthcare, academia and personal health. With regards to disorders of sleep, reliable and scalable sleep monitoring is set to provide a better understanding of sleep disorder progression and severity. This could facilitate better and earlier diagnosis and decision-making for individual patients, including in instances where individuals need to be progressed to a new treatment. Digitisation may also be used in disease prevention and to provide lifestyle recommendations. Objective ubiquitous monitoring of sleep-wake cycles, combined with multi-modal data inputs reflecting an individual's physical activity profiles, nutrition, all-day heart rate and genetic information will allow users to receive personalised feedback for health and well-being purposes and disease prevention. New technological advancements will allow for improved sleep coaching interventions that are aimed to improve sleep hygiene or provide with better recovery for example. Furthermore, data generated from these technologies could be used to help monitor the impact of pharmaceutical and post-operative interventions. Similarly, the accrued data gathered from clinical and epidemiological studies, as well as from commercial wearable devices, represents an unparalleled opportunity to deepen our understanding of the role of sleep in well-being and disease.

From the perspective of pharmaceutical companies, there are several benefits to the digitisation of sleep. Wearables offer the potential to deploy sleep monitoring at scale, in large populations that are required for late-phase clinical trials and can be used to provide better and earlier evidence of treatment efficacy in sleep disorders, thus facilitating the progression of promising candidates through trial phases. Further, there are implications for patient centricity. Across diseases, sleep is meaningful to patients and their health. It is therefore important to objectively assess sleep metrics such as sleep quality, WASO or time spent sleeping through quantitative measures, instead of relying on questionnaires. A summary of these metrics is provided in the Supplementary Figure 2.8. Low-burden monitoring will facilitate sleep collection in trials and potentially help to increase trial participation and reduce attrition. Moreover, many metrics of sleep are strongly tied to the quality of life, thus, industry may welcome the use of these sensing technologies for post-market surveillance. The added knowledge of a potential positive impact of medicine on patients' sleep quality may enable better reimbursement rates.

Ultimately, the digitisation of sleep could facilitate a truly personalised sleep monitoring experience, empowering people to improve their sleep [208]. However, the reproducibility and robustness of novel sleep monitoring and data analysis methods must be addressed prior to their use on large, longitudinal and multi-modal collaborative studies. The impact that these technologies can have on the management and understanding of sleep, as well as the treatment and prevention of sleep disorders, is set to be paradigm-changing. Industry, academic, public policy and clinical stakeholders should collaboratively enable this process of validation to take place, moving a step closer to truly personalised digital health.

In sum, digitisation of sleep and ubiquitous sleep monitoring will have important implications on the characterisation of sleep, diagnostics and therapeutics. Large-scale collection of objective, longitudinal sleep data through unobtrusive sleep sensing devices, as explored throughout this thesis, will facilitate epidemiological studies exploring the impact of sleep on health and disease. Furthermore, these applications will likely expand into sleep health, becoming increasingly accessible to individuals with the potential to empower and enable individuals to understand, manage and change their sleeping habits [153].

Supplementary Note 1

Sleep Metrics

Beyond the sleep staging guidelines provided by AASM, there are several sleep metrics that are commonly used when assessing sleep-wake cycles and are referred to throughout the first few chapters of this thesis. Sleep onset time is defined as the boundary that determines the transition between a period where the person is awake to when they are sleep. Similarly, the boundary between when a person is asleep and the transition to wake is known as the sleep awakening time. The following table introduces some of the most readily used sleep metrics based on these definitions which are used frequently used in the next few chapters.

Additionally, recently Phillips and colleagues described the Sleep Regularity Index (SRI) as "the likelihood that any two time-points, on a minute –to-minute basis, 24-hours apart were the same wake-sleep state, across all days [312]". So if we were to derive SRI using 30-second epochs on accelerometer data the SRI equation would be:

$$SRI = 100 + \frac{200}{M \cdot (N-1)} \sum_{j=1}^{M} \sum_{j=1}^{N-1} \sigma(S_{i,j}, S_{i+1,j})$$

Given N days of recorded divided into M (epoch=30s) daily epochs, suppose that $s_{i,j}$ = 1 if sleep on day i and epoch j and 0 if they are awake.

New sleep metrics based on EEG data are currently being derived, in search of establishing reliable sleep EEG biomarkers that could be used to phenotype patients [313, 314]. However, it is important that new sleep metrics also address the prevalence and representativeness of the data being used and account for it as well as the sampling and data-collection bias associated to sleep studies.

Supplementary Table 1

Table 2.2 **Conventional Sleep Metrics**

Metric	Formula
Sleep Period Duration	‖ Sleep Awakening Time – Sleep Onset Time ‖
Sleep Period	[Sleep Onset Time, Sleep Awakening Time]
Wake After Sleep Onset (WASO)	$\sum_{n=onset}^{awake}$ ‖ Wakefulness ‖
Sleep Latency	[Preceding Sedentary Time, Sleep Onset Time]
Total Time in Bed (mins)	‖ Sleep Awakening Time – Preceding Sedentary Time ‖
Total Sleep Time (mins)	‖ Sleep Period Duration - WASO - Latency ‖
Sleep Efficiency (SE)	Total Sleep Time / Total Minutes in Bed

Supplementary Note 2

Classification Metrics

Understanding how well a specific method is performing at the classification task is of great importance for research, clinical, industry and lifestyle applications. While evaluating a method or model's accuracy can be insightful, it is not sufficient. In sleep-wake and sleep stage classification task not all errors are equal. Indeed, there are two categories of error: predicting a negative when the instance is positive and predictive a positive when the instance is negative. Moreover, there are two categories of good prediction: successful prediction is termed true and unsuccessful prediction is termed false. These four variants form a confusion matrix:

Supplementary Table 2

Table 2.3 **Confusion matrix: understanding false positives and false negatives in classification tasks**

Data Class	Classified as pos	Classified as neg
pos	True positive (TP)	False negative (FN)
neg	False positive (FP)	True negative (TN)

From this, a variety of performance metrics can be derived, offering different perspectives on how the chosen method performs. Supplementary Table 3 introduces the most common metrics used for sleep-wake classification algorithm performance evaluation.

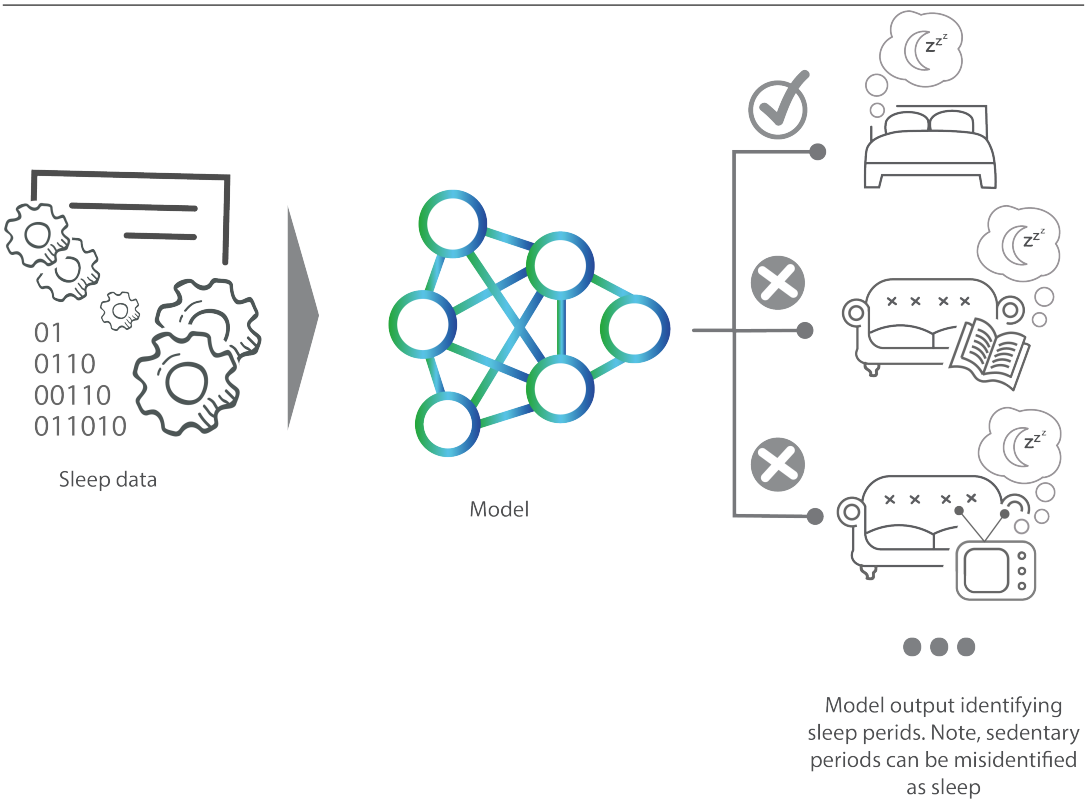
Supplementary Table 3

Table 2.4 **Assessing Classification Model Performance Through Metrics**

Measure	Formula	Function/ focus
Accuracy	$\frac{TP+TN}{TP+FN+TN+FP}$	Overall effectiveness for the algorithm
Precision (Positive Predictive Value)	$\frac{TP}{TP+FP}$	Agreement between the data labels and positive labels given by the algorithm
Recall (Sensitivity)	$\frac{TP}{TP+FN}$	Effectiveness of the algorithm to identify positive labels
Specificity	$\frac{TN}{TN+FP}$	Effectiveness of the algorithm to identify negative labels
F1 Score	$\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$	Conveys the balance between the precision and recall of the algorithm
AUC	$\frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$	Algorithm's ability to avoid false classification

Supplementary Figure 1

Figure 2.8 The Importance of Classification Metrics: Precision, Sensitivity and Specificity for Sleep Classification are paramount to understand if the model is not only accurate, but also capable of discerning sleep from sedentary behaviours or other bed activities.



Supplementary Note 3

Performance Metrics on Sleep-Wake Classification

From a classification perspective, the correctness of the classification algorithms can be evaluated by the number of instances or events correctly recognised class examples (true positives), the number of instances or events that are correctly identified as not belonging to a certain class (true negatives), as well as the instances or events that are wrongly classified as a given class (false positives) or were not recognised as a class (false negatives). Given these four metrics, we can compute what is known as a confusion matrix:

Based on the results obtained from the confusion matrix, there are several important metrics of performance that are derived to evaluate sleep classification algorithm performance. These are described in Supplementary Table 3.

Furthermore, when evaluating classification performance, other metric's are used depending on the characteristics of the data set. For instance, Cohen's Kappa is a metric that offers a comparison of the observed accuracy with respect to an expected accuracy (random chance, e.g. sleep-wake classification) and is defined as:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Where Pr(a) is the observed accuracy and Pr(e) is the expected accuracy. Other commonly used metrics of classification accuracy are Hamming and hinge loss, the Matthews correlation coefficient or zero-one classification loss.

Figure 2.8 exemplifies the classification performance of a classifier. There are several difference instances in which the participant may be sleeping or engaging in other, sedentary, activities (like reading or watching TV). The task of the classifier is to correctly classify sleep as such in blue and in purple as other waking activities (like sedentary behaviors). The sensitivity (or recall), specificity and precision are presented for this example classification outcome.

Supplementary Note 4

Actigraphy Specific Sleep Metrics

Actigraphy data can be analyzed and studied to focus in either sleep or circadian rhythms. Traditional sleep metrics like the ones explored previously (Wake after sleep onset, total sleep time, etc) can be extracted from actigraphy data. Moreover, metrics related to circadian rhythms can also be derived. Some of the most common ones are interday stability (IS), intraday variability (IV), amplitude at rest (L5) and relative amplitude (RA) [315]. These metrics provide information beyond that of traditional sleep metrics derived from actigraphy data. For instance, intraday variability (IV) is a measure of sleep fragmentation and interday stability (IS) can be used to asses sleep regularity.

CHAPTER 3

MULTIMODAL SLEEP STAGE CLASSIFICATION IN A LARGE, DIVERSE POPULATION USING MOVEMENT AND CARDIAC SENSING

Publications

Parts of this chapter have been published:

Zhai B.*, **Perez-Pozuelo I.***, Brage S., Guan Y. (2019). Ubiquitous monitoring of sleep-wake cycls using combined sensing and deep learning models. *In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE.*.

Zhai, B.*, **Perez-Pozuelo, I.***, Clifton, E. A., Palotti, J., Guan, Y. (2020). Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(2), 1-33.

* Equal contribution to this work

Contributions

I planned this project and designed the analysis plan in collaboration with my colleagues. I generated the scripts to process, model and visualise the data. I wrote this chapter and the resulting manuscripts.

3.1 Summary

Background

As explored in Chapter 2, wearable devices that collect movement and heart rate (HR) data increasingly provide a viable alternative to expensive, burdensome traditional methods of sleep monitoring, including polysomnography (PSG). However, few studies have combined actigraphy and HR data for sleep-wake and sleep-stage classification tasks. Further, as highlighted in the introductory chapter of this Thesis, the validity of wearable sensors at inferring sleep stages remains to be studied systematically.

Methods

Here, we present a set of algorithms for sleep-wake and sleep-stage classification based upon actigraphy and HR amongst 1,743 participants, with labels for training and model evaluation derived from PSG data. We present a processing pipeline, applicable to most research grade devices, deriving models and results in the largest open-access dataset for human sleep science.

Results

Our results demonstrated that neural network models outperform machine learning methods and heuristic models for both sleep-wake and sleep-stage classification. Convolutional neural networks (CNNs) and long-short term memory (LSTM) networks were the best performers for sleep-wake and sleep-stage classification, respectively. Using SHAP (SHapley Additive exPlanation) we identified that frequency features from cardiac sensors are critical to sleep-stage classification. Finally, we introduced an ensemble-based approach to sleep-stage classification, which outperformed all other benchmarks, achieving an accuracy of 78.2% and F_1 score of 69.6% on the classification task for three sleep stages.

Conclusions

Together, this work represents the first systematic multimodal evaluation of sleep-wake and sleep-stage classification in a large, diverse population. Alongside the presentation of a novel, accurate sleep-stage classification approach, the results highlight multimodal wearable sensing approaches as scalable alternatives to PSG for accurate sleep-classification, providing algorithms for classification tasks and guidance on optimal algorithm deployment based upon the task.

3.2 Background

Sleep is a reversible physiological state that is essential for life, health and performance. Whilst the functions of sleep are not yet fully understood, it is known to restore energy, promote healing, rejuvenate physical systems, interact with the immune system and influence brain function, with manifold consequences including for memory consolidation and behaviour [316–319]. As a result of its importance to vital human processes and the incomplete understanding of its function, accurate sleep monitoring is of interest to the understanding of human health and an active area of research for the *ubiquitous computing* community [320].

Chapter Significance: This chapter provides a holistic overview of the strengths and limitations of multimodal (movement and heart rate) wearable sensors on sleep-wake and sleep-stage classification. The strength of this work lies on the diversity and size of the population used, the systematic nature of the experiments conducted and the explainability section of these approaches. Overall, the chapter highlights that while modern wearable devices using PPG and acceleration can do a fairly good job at classifying sleep-wake and NREM, REM and wake, when the level of granularity is higher than that, the performance of these devices is compromised.

Traditionally, human sleep has been monitored in laboratory settings using polysomnography (PSG). PSG is a multi-sensor approach to monitoring, involving the collection and conveyance of a range of different signals from many sensors operating simultaneously. These sensors include electroencephalography (EEG), electromyography (EMG) and electrooculography (EOG), which together facilitate the measurement of brain activity, alongside both muscle and eye movement. Measurements of respiratory and cardiac activity are also often included. PSG recordings are processed and segmented into epochs, typically of 30-second duration, and an expert or technician then assigns a sleep stage to each epoch. This sleep stage "scoring" typically follows the rules set out by the American Association of Sleep Medicine (AASM), which defines five stages of sleep-wake cycles: wake (W), rapid eye movement (REM) sleep and three types of non-REM sleep (NREM), known as N1, N2 and N3 [321].

Whilst traditional PSG is considered the gold-standard for sleep monitoring, as a result of the need for sensing equipment, its use is limited to laboratory settings and typically to just one or two nights. These single nights of observed sleep in an unfamiliar environment may not reflect normal sleep. Further, it is impractical to measure sleep using this method for more than two consecutive nights as it is burdensome to patients or study participants. PSG is also expensive and requires expert set-up and analysis. For these reasons, efforts to monitor individual's typical sleep duration and quality longitudinally in large, free-living populations have generally relied upon sleep diaries or self-reported questionnaire data. Whilst sleep diaries are cost-effective, scalable and able to collect information regarding typical sleep patterns,

there are concerns as to the validity and reliability of participant responses [322]. Wearable sensors offer a potential solution. Such sensors provide valuable, unobtrusive tools through which to objectively monitor physical activity in large population studies studies, with potential applications for sleep monitoring.

Conventional approaches to monitoring sleep using wearable devices are primarily based on actigraphy (count-based movement information) and accelerometry (raw, high frequency data which is often triaxial) [81, 323, 82, 324]. However, recent technological and battery life advances increasingly facilitate multimodal sensing (i.e., combining accelerometry with HR sensing). Multimodal sensing facilitates more intricate human activity recognition (HAR) tasks and has shown promise for sleep-stage classification [325]. The validity of actigraphy for the classification of sleep-wake transitions has been demonstrated over the past three decades [81, 324, 323, 82]. Algorithms applied to actigraphy for this purpose exploit differences in body movement between wakefulness and sleep. Recent work has demonstrated how different methods for binary sleep-wake classification using actigraphy compare when applied to the same, standardized dataset [237]. Furthermore, HR variability (HRV) metrics could be valuable for multistage classification as autonomic function fluctuations occur between non-REM sleep and waking/REM sleep, whilst these same functions are consistent when comparing wake to REM [326–329].

Understanding time spent in different sleep stages (beyond binary sleep-wake classification) in free-living environments has important implications for commercial applications, as well as for research. For example, accurate sleep architecture inferences may provide better information to guide sleep-related behavioural changes and recommendations [330]. To date, sleep stage assessment could only be achieved using PSG, which cannot be applied to large, population-based studies of free-living individuals with the power to make inferences regarding the implications of sleep for health and illness. There exists very limited literature regarding the performance of multimodal sensing using wearable technologies in sleep-wake classification or sleep-stage classification, and the methods used for these tasks [331]. Given the increasing popularity of these technologies for both commercial and research applications, the development of a set of benchmarks to evaluate the performance of sleep-wake and sleep-stage classification methods on multimodal wearable data using actigraphy and HR sensing would address a major gap in the existing literature.

Remark: In order to address this gap, this work focused on five major contributions:

1. We introduce a framework for pre-processing and analyzing multimodal sensor data from movement (actigraphy) and cardiac (RR intervals from ECG) sensors. We use the same signals that are derivable from research-grade ECG or photoplethysmogram (PPG) devices.

2. We systematically compare single modality sensors to combined sensing (actigraphy + HR/HRV) approaches for classifying sleep-wake using different machine learning models.

3. We extended this systematic comparison to explore the performance of single modality approaches and combined sensors across three different multistage classification tasks: (A) Task 2, conventional three-stage classification(NREM, REM, Wake), (B) Task 3, four-stage classification (light sleep, deep sleep, REM and wake) and (C) Task 4, five-stage classification (AASM-standard), also using benchmark machine learning (ML) and deep learning (DL) models.

4. We introduce time deviation per class, a new evaluation metric that represents an easy-to-interpret measurement of the predicted results from a healthcare practitioner perspective. This complements conventional machine learning metrics, including confusion matrices. Furthermore, we explore model explainability and individual sensor contributions using SHAP, a unified approach to explain the output of tree model based ML and DL models.

5. We introduce an ensemble structure for multistage sleep classification based on multi-timescale and multimodal DL *ensemble architecture*. This architecture aims to address the needs encountered in free-living environments whilst exploiting the individual contributions and strengths of different classifiers.

To our knowledge, our work presents the first systematic multimodal and multistage evaluation of sleep-wake cycles and sleep stages in a large, diverse population. We examine each individual method and modality, as well as exploring how sensor fusion leads to better performance. Additionally, we explore the features that contribute the most to the different classification tasks, developing an understanding of the physiological underpinnings of our models.

3.3 Related work

Since the 1980s, a vast number of studies have explored new methods and techniques to infer sleep-wake cycles using actigraphy with either single-axial [82, 81, 324] or, more recently, tri-axial accelerometry [332]. While these methods have proven valuable, they were often derived in small cohorts. The recent availability of large datasets, provided by initiatives such as the National Sleep Research Resource [333, 334][1], makes it possible for researchers to create large standardized benchmarks, such as that proposed in our work, for the first time. For example, Palotti et al. leveraged one of the available datasets, the Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Study[2], to compare the performance of the most relevant heuristic approaches and ML methods for binary sleep-wake classification [237]. Whilst novel, their work was limited by: (1) exclusively comparing methods for sleep-wake classification, rather than multistage classification; (2) only using actigraphy data. Here we address these limitations in the MESA Sleep Study dataset, the only dataset suitable for such experiments to-date.

Beyond actigraphy data, in this work, we explore the use of HR and HRV data. HR can be defined as the average number of heartbeats per minute, while HRV is a measure of the variability in beat-to-beat intervals, known as RR intervals. These measurements are powerful biomarkers that have been used to understand training and recovery, address chronic disease and monitor stress and sleep [335–337]. HRV is typically higher during the night, reflecting the fact that sleep is a state in which vagal activity, characterized by rapid fluctuations in activity controlling coronary artery tone, HR and systolic blood pressure, is dominant [335–337]. Thus, HRV shows a nocturnal increase in the deviation of mean RR intervals. These deviations also differ between sleep stages. Conversely, several studies have shown that HR does not change significantly between sleep stages, although some work has suggested a rise during REM sleep [338, 339]. HRV analysis has demonstrated that the High Frequency (HF) band doubles in relative power when going from quiet wakefulness to non-REM sleep [328]. Hence, a full-feature set of HRV-relevant features is a powerful tool for sleep-stage classification [340]. Recently, Radha et al. have reported that HRV has great potential to classify sleep stages [341]. However, their work was performed in a private dataset and conducted using some features that are often not present on wearable devices. In our work, we devise HR/HRV features that could be extracted from research-grade wearable sensors and evaluate their performance in the largest public dataset to date.

Beyond ML and DL models, ensemble architectures are becoming increasingly prevalent for HAR tasks. For instance, in 2015, Single et al. adopted an approach consisting of three variants of long-short term memory (LSTM) networks that worked in parallel to tackle a biological sequence analysis task and then used *majority voting* to decide upon the final classification

[1] https://sleepdata.org
[2] https://sleepdata.org/datasets/mesa

prediction [342]. In [343], Guan and Ploetz developed a LSTM ensemble model via epoch-wise bagging for efficient training. They injected several random factors to increase the diversity of the classifiers and improve performance in several HAR tasks.

Recent work has explored the application of ensemble models for automatic sleep classification using PSG/EEG signals. These studies have shown promising results, improving the performance of shallow ML and even DL approaches [344–346]. Koley et al. used an ML ensemble architecture approach consisting of five binary support vector machine (SVM) classifiers to classify different sleep stages [344]. Using a *"winner-takes-all"* ensemble method [347] the researchers achieved significantly better discriminant capability relative to single SVM models applied to EEG signals. Recently, Huy et al. applied an ensemble method to multimodal PSG data (EOG, EEG and electromyography (EMG)) by fusing classifiers [348]. All these previously reported models are based on sleep epoch (30 seconds) level feature extraction protocols and use classifier ensembles derived in sensors which often exceeded 100Hz sampling rates (EEG data, etc). These methods demand high specifications with regards to computing power. Currently, the limited data storage and processing capabilities of wearable devices mean that they are unlikely to be able to support these models in free-living conditions.

3.4 Methods

The MESA Sleep dataset is introduced and described in Section 3.4.1 [333, 334]. All experiments reported here were conducted based in this dataset. In Section 3.4.2, we provide an overview of the data pre-processing and feature extraction method for modalities which consist of cardiac sensing (HR and HRV) and movement sensing (actigraphy). All tasks explored, including our *ensemble method*, are introduced in Section 3.4.3. In Section 3.4.4, we introduce the models used for our benchmark work. Section 3.4.5 describes how we designed our experiments. Finally, in Section 3.4.6, the metrics used to evaluate the classification models are discussed.

3.4.1 Dataset Description

The Multi-Ethnic Study of Atherosclerosis (MESA) dataset is a multi-centre longitudinal study designed to investigate the characteristics of sub-clinical cardiac disease. The study comprises 6814 asymptomatic men and women of black, white, Hispanic and Chinese-American ethnicity, of which 2,237 were also enrolled in the MESA Sleep Study. As part of the MESA Sleep Study, all participants wore an actigraphy device for one week and underwent concurrent PSG for one night. Data for this study was acquired in six different centers across the US and followed the appropriate Institutional Review Board approvals and written informed consent for participant data acquisition [333, 334].

The MESA Sleep Study was conducted using a Compumedics Somte System for PSG, which includes the ECG signals here used to derive HR and HRV and their associated features, alongside an Actiwatch Spectrum from Philips Respironics to record actigraphy data. This device captures measurements of movements defined as "activity counts"[3] and aggregates them into 30 second epochs. The Actiwatch was securely fastened to participant's non-dominant wrist. These actigraphy signals and their associated features can be derived in most research-grade wearable devices. The sensors for the Compumedics PSG comprised: cortical EEG, bilateral EOG, chin EMG, abdominal and thoracic respiratory inductance plethysmography, airflow, ECG, leg movement sensor and finger pulse oximetry. These sensors collected three types of signals: bioelectrical potentials (EEG, EOG, EMG, ECG), waveforms received from transducers (thermistors on the airflow devices, inductance respiratory bands, piezo leg sensors and position sensors from the leg device) and auxiliary devices (oximetry measures of oxyhemoglobin saturation and nasal pressure records). Full details of the setup, protocol and sampling rates are available [4,5]. All participants included in our study had at least one full night of PSG recording with concurrent actigraphy and ECG. An illustration of the experimental set up is provided in Figure 3.1. All nocturnal recordings were transmitted to a centralized reading

[3] https://www.salusa.se/Filer/Produktinfo/Aktivitet/TheActiwatchUserManualV7.2.pdf
[4] https://sleepdata.org/datasets/mesa/pages/equipment/montage-and-sampling-rate-information.md
[5] https://sleepdata.org/datasets/mesa/files/documentation

Figure 3.1 Experimental setup and tasks: Our models are trained using a combined-sensing, multimodal approach which incorporates two time-series signals: actigraphy and ECG derived HR and HRV and uses Gold-Standard PSG for the training labels

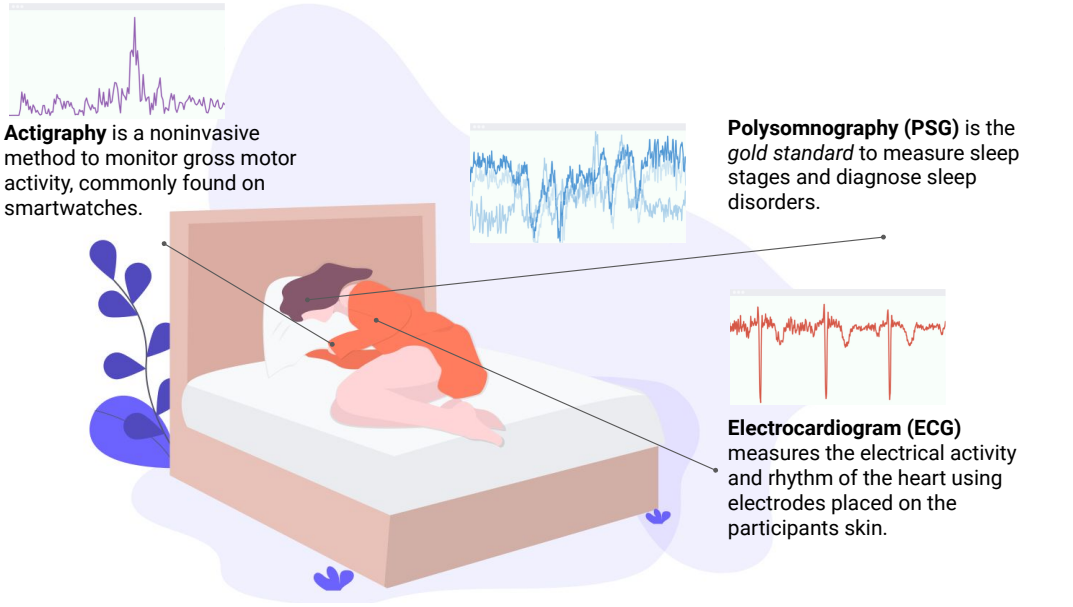


Table 3.1 Breakdown of population based on sex, age and demographic characteristics, by dataset (training or test).

Dataset	Total	Female	Male	Black	Chinese-American	White	Hispanic	Age ($\mu \pm \sigma$)	Min Age	Max Age
Training	1395	752 (54%)	643 (46%)	383 (28%)	153 (11%)	511 (37%)	348 (25%)	69.29 ± 8.73	54	94
Test	348	198 (57%)	150 (43%)	103 (30%)	39 (11%)	128 (37%)	78 (22%)	68.52 ± 9.21	55	89

Numbers are N, N(%) or mean (SD). Age is given in years.

Table 3.2 **Sleep statistics of participants in the study.**

Dataset	Total Sleep Time (TST)	Total Time in Bed (TIB)	Sleep Efficiency (%)	Wake After Sleep Onset (WASO)	N1	N2	N3	REM
All	359.0±80.7	475.0±85.3	76±13.1	90.3±62.7	49.42±30.9	207.0±60.1	40.0±33.4	66.9±28.9
Training	357.5±80.2	473.5±84.9	75.9±13.1	90.5±62.7	49.3±31.1	206.6±60.4	39.6±33.3	66.6±29.3
Test	365.1±82.5	480.9±87.0	76.4±12.8	89.2±62.6	49.8±30.4	208.5±58.9	41.68±33.7	68.3±27.4

Numbers are minutes except sleep efficiency measured in percentage(mean ± SD)

center at the Brigham and Women's Hospital (Boston, MA, USA) and data was scored by trained technicians using AASM guidelines. For our training labels, we used the expert scoring and epoch staging annotations on PSG data provided by Bild et al [349]. Note that the MESA Sleep dataset is the **only large open-access dataset** combining gold-standard measures of sleep through PSG with wearable sensor data from actigraphy as well as ECG (HR/HRV) and thus the only existing dataset appropriate for our purposes.

Table 3.1 summarizes the main demographic characteristics of the participants by training and test splits.

3.4.2 Data Pre-processing and Feature Extraction

In this work, we synchronized PSG, ECG and actigraphy records into 30-second sleep epochs for 1,743 of the 2,237 participants included in the study. A total of 494 participants were excluded on the basis of: (1) lack of concurrent PSG, ECG and actigraphy data; (2) lack of enough quality standard data ($<$ 1.5 h of usable data from the concurrent three sensing methods); or (3) lack of data integrity or misalignment of data, we removed actigraphy outlier epochs based on human expert annotations. These outliers are either non-wearing periods or equipment failure periods. For actigraphy epochs labeled as outliers, their corresponding HR/HRV epochs were also removed [350].

For generalisability purposes, we included diseased participants in our analysis; full details are presented in Supplementary Table S1. Similarly, we did not exclude a total of 30 subjects (about 2% of the total cohort) who do not have any REM epochs at all, although we do understand that these sleep patterns are physiologically very unlikely. The sleep stages for subjects in this dataset were scored by individual sleep technicians, blind to the disease status of the participants, into five classes (wake, N1, N2, N3, REM) according to AASM guidelines [349].

For the ECG signal, we derived features that are only based on RR intervals instead of using the raw ECG signal. The rationale behind this was to make our work as transferable as possible to data collected from research-grade devices such as miniaturized ECGs or wrist wearables that incorporate PPG sensors (i.e., the Empatica E4 wristband). Participants whose ECG records did not include a full night of sleep, or whose data was corrupted were excluded from further analysis.

QRS complexes (R-points) were detected using Compumedics Somte (Abbotsford, VIC, Australia) software Version 2.10 (Builds 99 to 101). The R-points were classified as normal sinus, supraventricular premature complex or ventricular premature complex. Data cleaning, filtering and noise removal took place during this step of the process using the Python package HRV-analysis[6]. First, RR interval outlier data was filtered using a threshold method with a range between 300 to 2000 ms following the method previously described by Tanaka et al. [351], then the ectopic beats were removed by through the methods described in Malik et al. [352]. Second, we linearly interpolated the removed R-points. We grouped the RR intervals into 30 seconds to match the time interval of actigraphy data. Recalling the description in Section **??**, the HRV describes the physiological variation of the beat-to-beat interval that can be extracted from the time-distance between adjacent R wave peaks. Thus, we calculated 30 cardiac features from each 30-second window length that matches the epoch of the actigraphy data. Following the approach used by Radha & Fonseca et al. [341], we extracted features in four domains (time, geometrical, frequency and non-linear domains). Table 3.4 details the full set of cardiac features used in this work.

[6] https://pypi.org/project/hrv-analysis/

Table 3.3 **Full set of features extracted from the actigraphy signal.**

Feature name	Description
Activity Count *	*Raw activity count from the actigraphy device*
Log Activity Count *	*Natural Logarithm of the activity count*
Mean Activity *	*Mean value for the window of activity of size N.* $1 \leq N < 20$
Median Activity *	*Median value for the window of activity of size N.* $1 \leq N < 20$
Std Activity *	*Standard deviation value for the window of activity of size N.* $1 \leq N < 20$
Variance Activity *	*Variance value for the window of activity of size N.* $1 \leq N < 20$
Minimum Activity *	*Minimum value for the window of activity of size N.* $1 \leq N < 20$
Maximum Activity *	*Maximum value for the window of activity of size N.* $1 \leq N < 20$
NAT Activity *	*Number of epochs, in a window of size N, which the value for the activity count is larger than 50 and lower than 100. Devised from [81].* $1 \leq N < 20$
Any Activity *	*Number of epochs that contain any activity in the window of size N.* $1 \leq N < 20$
Skewness of Activity *	*Skewness for the window of activity of size N.* $4 \leq N < 20$
Kurtosis of Activity *	*Kurtosis for the window of activity of size N.* $4 \leq N < 20$

We adopted two strategies for extracting actigraphy-related features. For DL approaches, that can automatically extract features from the raw inputs, the raw activity counts (sampling rate at 1/30 Hz) were extracted from the device as the input. For the other ML models, a total of 370 handcrafted time-series features were extracted, as described below. These features have been commonly used in the literature (i.e., [353, 354, 237]).

For each sleep epoch T, we calculated summary statistics (i.e., mean, variance, median, kurtosis) for actigraphy data that considering both centered and non-centered sliding windows of N sleep epochs (with $N = \{1, 2, ..., 19\}$), where each sleep epoch contain a scalar value. We also calculated other commonly used metrics, such as the raw and natural logarithm values of the activity counts for each epoch T. These features are listed in Table 3.3.

Our full feature set (i.e., both activity and cardiac features) were normalized using the z-score method. A summary of the pipeline used in this work is shown in Figure 3.2.

3.4.3 Tasks

We structured our work on 5 different tasks that allowed us to explore the objectives of this work and test several hypotheses based on multimodal fusion and new model development.

Our first task, **Task 1**, aims to establish benchmarks for sleep-wake (binary) classification using single modality (either actigraphy or HR/HRV) and multimodality approaches (combining both modalities). In doing so, we compare conventional statistical learning methods and simple neural network methods across modalities. This task is the most explored one among the research community in this area [82, 81, 323, 354, 356] and we also aimed to augment the benchmarks previously reported by Palotti et al. [237].

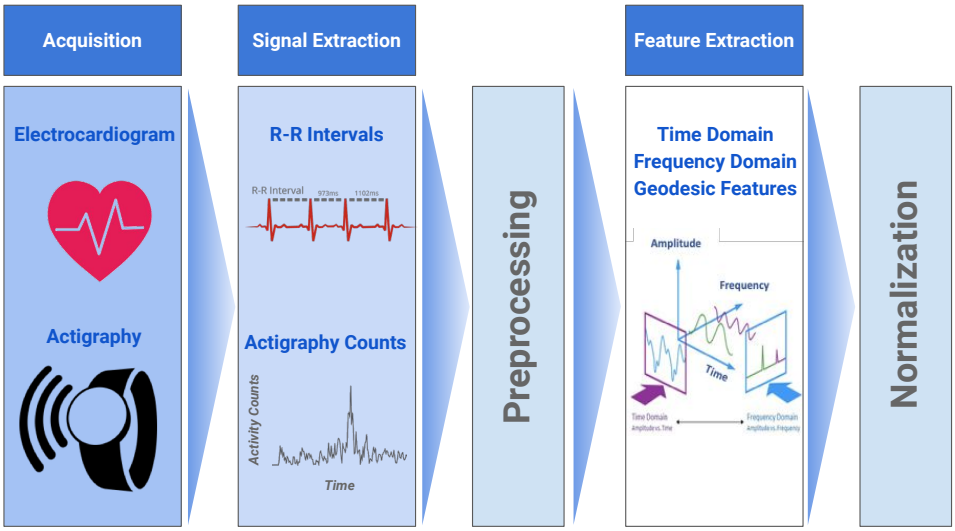
Table 3.4 **Full set of cardiovascular related features grouped by domain.**

Time Domain Features	
Mean HR ♥	*Mean heart rate for that window*
Maximum HR ♥	*Maximum heart rate for that window*
Minimum HR ♥	*Minimum heart rate for that window*
Std HR ♥	*Standard deviation for the heart rate for that window*
SDNN ♥	*Standard deviation of Normal-to-Normal interval (NNi)*
SDSD ♥	*Standard deviation of NNi differences*
NN50 ♥	*Number of NNi differences greater 50ms*
pNN50 ♥	*Ratio between NN50 and total number of NNi*
NN20 ♥	*Number of NNi differences greater 20ms*
pNN20 ♥	*Ratio between NN20 and total number of NNi*
RMSSD ♥	*Root mean of squared NNi differences*
Median NNi ♥	*Median of NNis*
Range NNi ♥	*Range between smallest RR intervals to largest RR intervals*
CVSD ♥	The coefficient of variation of successive differences , the RMSSD divided by mean NNi
Coeff. of Variation of NNI ♥	*The Coefficient of Variation of NNi, i.e. the ratio of sdNN divided by mean NNi*
Geometrical Domain Features	
Triangular Index ♥	*The HRV triangular index measurement is the integral of the density distribution (that is, the number of all RR intervals) divided by the maximum of the density distribution (class width of 8ms)*
Frequency Domain Features	
Low Frequency ♥	*Low Frequency is the variance (i.e., power) in HRV in the Low Frequency (.04 to .15 Hz). Reflects a mixture of sympathetic and parasympathetic activity*
High Frequency ♥	*High Frequency is the variance (i.e., power) in HRV in the High Frequency (.15 to .40 Hz). Reflects fast changes in beat-to-beat variability due to parasympathetic (vagal) activity*
Variance in Low Freq. ♥	*VLF is the variance (i.e., power) in HRV in the Very Low Frequency (.003 to .04 Hz). Reflect an intrinsic rhythm produced by the heart which is modulated by primarily by sympathetic activity*
Low/High Freq. Ratio ♥	*The LF/HF ratio is sometimes used by some investigators as a quantitative mirror of the sympathy/vagal balance*
Norm. Low Freq. Ratio ♥	*Normalized low frequency ratio calculated from the raw values of low frequency band (LF or HF) divided by the total spectral power*
Norm. High Freq. Ratio ♥	*Normalized high frequency ratio calculated from the raw values of high frequency band (LF or HF) divided by the total spectral power*
Mean NNi ♥	*Mean over the RR intervals*
Total Power ♥	*Total power of the density spectral*
Non-Linear Domain Features	
Cardiac Sympathetic IdNx ♥	*Cardiac Sympathetic Index [355]*
Mod. Cardiac Symp. IdNx ♥	*A modified cardiac sympathetic index calculated by* $\frac{SD2^2}{SD1}$
Cardiac Vagal IndeNx ♥	*Cardiac Vagal IndeNx [355]*
SD1 ♥	*Poincaré plot standard deviation perpendicular the line of identity*
SD2 ♥	*Poincaré plot standard deviation along the line of identity*
SD1/SD2 Ratio ♥	*Ratio of SD1 to SD2*

Task 2 consisted of the same systematic evaluation, but this time, the simplest sleep staging paradigm was introduced (Wake, NREM, REM). Here, the AASM scores provided in the MESA dataset are simplified and collapsed into a simpler representation of sleep staging. Wake and REM remain the same, but N1, N2 and N3 are grouped together to become NREM sleep as an entity. The feasibility of this task has also been tested by other studies [341, 331].

Taking a step further in the level of granularity, **Task 3** classifies the data into Wake, REM, Light Sleep and Deep Sleep. Here, light sleep captured both N1 and N2 which is often considered a transition state between light and deep sleep and usually takes up the largest percentage of time during a full sleep cycle [357]. Given the heterogeneity and prevalence of

Figure 3.2 Multimodal data processing pipeline: after removing low quality data, the signals from the actigraphy device and ECG are synchronized and features are extracted and normalized.



N2, the difficulty of the task has risen significantly. The models are expected to perform worse than they did on previous tasks.

Task 4 explored the classification of sleep stages based on AASM rules (Wake, REM, N1, N2, N3). This task has the highest level of granularity, and it is, in fact, a task in which even the current state of the art, DL approaches on gold-standard PSG recordings often do not achieve satisfactory performance [346]. This task faces two challenges. The first being the class imbalance, as N1 and N3 sleep epochs account for only 11% and 7% of the data respectively. The second challenge is the nature of our modalities that do not capture direct cortical signals, compromising the performance in more granular classification tasks.

Finally, we introduced an *ensemble method* which aims to combine the unique *perspectives* and *capabilities* of DL classifiers with different window sizes containing discriminant power from different temporal dependencies that could be characteristic of different sleep stages.

3.4.4 Models and settings

Conventional heuristic approaches have been readily used in the past 30 years for Task 1 (binary sleep-wake classification). It has recently been shown that feature-based ML and DL approaches greatly outperform all these methods [237].

ML and DL techniques are increasingly used in medical sciences [346, 331]. Here, we use supervised learning techniques on time-series data. This entails generating models that learn

Table 3.5 **Experiment settings based on input modalities**, where l is the window length of the input ($l = \{20, 50, 100\}$), the inputs are for each sleep epoch

Algorithm Type	Modality	Input Dimension	Features Used (Full list on Tables 3.3 and 3.4)
ML	*[Actigraphy]*	$x \in \mathbb{R}^{370}$	370 features were derived from Activity Counts
	[HR/HRV]	$x \in \mathbb{R}^{30}$	30 features were derived from RR intervals
	[HR/HRV, Actigraphy]	$x \in \mathbb{R}^{400}$	Concatenation of the two modalities above
DL	*[Actigraphy]*	$x \in \mathbb{R}^{l}$	Activity Counts
	[HR/HRV]	$X \in \mathbb{R}^{l \times 8}$	8 features were derived from RR intervals: Mean NNi, Standard Derivation of RR interval (SDNN), RR interval differences (SDSD), Very Low Frequency, Low Frequency, High Frequency Bands, Low Frequency to High Frequency Ratio and Total Power.
	[HR/HRV, Actigraphy]	$X \in \mathbb{R}^{l \times 9}$	Concatenation of the two modalities above

mappings between input and output spaces. For instance, Random Forest (RF) approaches have shown strong performance on activity recognition tasks [358]. Similarly, Radu et al. [46] showed promising results using DL approaches on multimodal sensor data for activity and context recognition tasks. Indeed, wearable sensors exploiting multimodal approaches have shown the advantages of this methods over single modality approaches for human activity recognition tasks [359]. Going beyond traditional activity recognition tasks, ML and DL models have been shown to outperform conventional heuristic approaches for actigraphy sleep-wake classification [237, 353]. DL models have also shown great promise in the automatic classification of sleep stages using EEG or multimodal sensor data [348]. Here, we expand that work to multimodal wearable and minimally obtrusive sensors by systematically evaluating how the most well-established ML and DL models perform based on different sensor modalities and when using combined sensing.

For all included tasks and modalities, we explored the most common shallow ML and DL architectures. These comprise linear support vector machines, logistic regression, random forest, perceptrons, convolutional neural networks (CNN) and long-short term memory networks (LSTM). We hypothesised that given the large amounts of data DL models would be better suited. Details on the ML and DL classifier settings can be found in Table **??**.

3.4.5 Experimental Design

Once the feature sets were built for our two input streams, we randomly split the dataset into training and test sets following an 80/20 split where 80% (1,395 subjects) went to the training set and 20% (348 subjects) went to the test set. More details including demographic information can be found in Table 3.1 and a summary of sleep statistics is introduced in Table 3.2. .

The inputs to our single modality and multimodal experiments can be found in Table 3.5. When using multimodal approaches, we used a *channel-wise* stacking approach prior to inputting the

resulting matrix into our models. These methods were adopted across all benchmarks for our tasks. Following the method used in [237], our hyperparameter search is described below:

- **ML hyperparameter search:** we employed 5-fold cross-validation on the training set.
- **DL hyperparameter search:** we employed a hold-out method to randomly split the training dataset into a validation set of 279 subjects (20%) and a training set of 1,116 (80%).

The full detailed list for our hyperparameter tuning can be found in Table **??**. We used Scikit-learn[7], Keras[8] and Tensorflow[9] to implement our models. For our feature set, we emulated previously used approaches [341, 237] for movement and cardiac sensor feature extraction in traditional ML and DL setups, which were mentioned in section 3.2. In our ML experiments, each input vector contains 400 features that combined 370 statistical features extracted from actigraphy and 30 HR/HRV features as we describe in the feature engineering section. As such, the single modality approaches for actigraphy input 370 features whereas for HR/HRV 30 features for each sleep epoch were included for each input vector. These were used as inputs for our feature-based ML benchmarks.

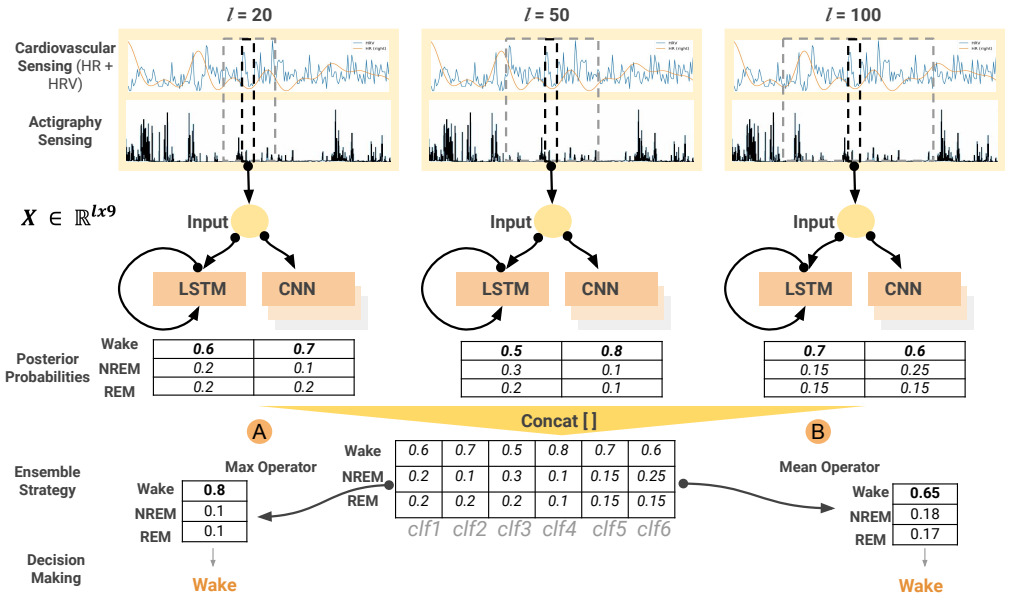
In our DL experiments, given that we wanted our work to be as transferable and device-agnostic as possible, we decided not to use the raw ECG signal. ECG signals are expensive and are not currently available in most of the wearable sensors. Thus, instead, we used an 8-dimensional HR/HRV feature set (see Table 3.5) that can be derived from many wearable cardiac sensors, such as Epatica[10], or ActiHeart[11]. For movement data, we simply use the activity counts that can be acquired directly from the wrist-worn actigraphy device.

Due to the fact that the DL techniques used CNN and LSTM are able to learn high-level features from the data, their input is a window of a size which can be either 20, 50 or 100 sleep epochs in width. Similarly, the conceptualization of the proposed *ensemble model* is based on how domain experts, in this case physicians and laboratory technicians who examine PSG records, approach multistage sleep classification.

For PSG annotation, sleep technicians and physicians often use *adjacent* information as well as contextual temporal information to inform their decisions in scoring a particular epoch. Indeed, they may look at information and trends within a 30 minute or 1 hour period as well as contextual information regarding the distribution of previous sleep stages to reach a decision for their scoring.

[7] https://scikit-learn.org
[8] https://keras.io
[9] https://tensorflow.org
[10] https://www.empatica.com
[11] www.camntech.com

Figure 3.3 Ensemble model: The model starts by taking inputs from different window lengths (l) from the multimodal sensors. A total of six different classifiers are used, combining a mixture of CNNs and LSTMs and exploiting their individual strengths. This results on posterior probability confusion matrix that is then combined through concatenation as part of the ensemble architecture. Finally, the decision making layer takes place by either (a) using a maximum operator approach or (b) a mean operator across all classifiers



Inspired by this procedure that expert PSG scorers undergo when analyzing laboratory datasets, we use an *ensemble classifier model* to create an "aggregation" of individual learned models which are trained using different *temporal perspectives* of the data. By doing so, individual learners *focus* on different aspects of a particular sequence, which is to lead to a more robust, and typically, a better outcome for recognition performance when using aggregated results. Our *ensemble model* combines the best 6 classifiers (a combination of CNNs and LSTMs) with various sliding window lengths (20, 50 and 100 sleep epochs, their corresponding sleep recording length of 10, 25 and 50 minutes). This approach aims to offer low bias and low variance due to it's capabilities in capturing various temporal dependencies. Following the previously described ensemble model pipelines, we explore two score-level fusion methods for variance reduction. These methods are *model-averaging* and *maximum posterior selection*.

The sleep stage ensemble classification model is based on standard single-layer CNN and LSTM networks, as summarized in the previous four tasks. Figure 3.3 illustrates the structure for individual classifiers and their score level fusion mechanism. At the training stage, each classifier is trained independently given the hypothesis that data from sliding windows of different lengths carry different discriminative information for each sleep-stage class.

After the forward propagation takes place for a sleep epoch, each classifier's softmax layer produces a K-dimensional probability vector $\mathbf{p}_t^m \in \mathbb{R}^K$, where K is the number of classes in a particular task and t is the sleep epoch index(for instance,a 450 minutes sleep recording has

900 indexes , so $t = \{1, 2, ..., 900\}$), m is the m^{th} DL classifiers in our ensemble model and we have 6 classifers so $m = \{1, 2, ..., 6\}$. Probability vectors from all $M = 6$ models are then combined into a probability matrix (each mode is independent of each other). Then, the two different score-level fusion methods are applied on these posteriors. The first method takes their arithmetic mean resulting in a final score vector $\mathbf{p}_t^{fusion} \in \mathbb{R}^K$:

$$\mathbf{p}_t^{fusion} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{p}_t^m$$

and then a class($\hat{k}_t$ represents the predicted sleep stage) is assigned to the one with the highest probability:

$$\hat{k}_t = \arg\max \mathbf{p}_t^{fusion},$$

The second method is to take the maximum posterior probability across all classifiers for a certain class k, and then the class label $\hat{k}_t$ is assigned to the one with highest probability which is the k^{th} class such that:

$$\hat{k}_t = \arg\max_k \mathbf{P}_t, \mathbf{P}_t = [\mathbf{p}_t^1, \mathbf{p}_t^2, ...\mathbf{p}_t^m]$$

where $\mathbf{p}_t^m$ is the posterior probability of the m^{th} classifier. This method will firstly find the posterior probability vector that contains the highest value, then finds the the maximum value's index k which refers to the k^{th} class. And t refers to the sleep epoch index. The whole ensemble scheme is also shown in Figure 3.3.

3.4.6 Evaluation Metrics

We adopted commonly used metrics in machine learning and medical sciences to evaluate the performance of the different classification algorithms based on task and modality combination. All of our performance metrics were derived at both a *subject level* (first derived on an individual by individual basis and then averaged across the population) and a *group level*.

To assess class imbalance and evaluate performance, we adopt several popular metrics based on True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) classifications. These evaluation metrics can be summarized as follows:

- **Accuracy** counts the number of correctly classified sleep epochs, normalized over the total number of sleep epochs (Acc $= \frac{TP+TN}{TP+TN+FP+FN}$).

- **Recall** measures the proportion of positives that are correctly identified as the given stage ($R = \frac{TP}{TP+FN}$).
- **Specificity**, also known as true negative rate, measures the proportion of negatives that are correctly identified as the given stage ($S = \frac{TN}{FP+TN}$).
- **Precision** is the fraction of correct classified instances among the overall positive predictions ($P = \frac{TP}{TP+FP}$).
- **$\mathbf{F_1}$ score ($\mathbf{F_1}$)** conveys the balance, with the harmonic mean, between precision and recall ($F_1 = 2 \times \frac{\mathrm{P}\times\mathrm{R}}{P+R}$).
- **Cohen's Kappa (κ)** measures inter-rater reliability/agreement, comparing observed accuracy with an expected accuracy ($\kappa = \frac{P_o - P_e}{1-P_e}$, where P_o the observed proportional agreement and and P_e the expected proportion of agreement). In this context, Cohen's κ factors out agreement by chance arising from the class imbalance of different sleep stages throughout the night.

With the exception of Task 1 (binary sleep-wake classification), all other tasks were multi-class classification and we average the results with *macro* average given the consideration of imbalanced data for the classification task.For all of our classifiers, we also computed confusion matrices to understand further where the different approaches excelled and where they fell short. We computed Cohen's Kappa to evaluate agreement across the whole population. Moreover, we used two-tailed t-tests with a significance level of 95% (significant differences when $p < 0.05$).

We also propose in this work a measure to intuitively understand how long, in terms of minutes, a classifier is under/over estimating sleep in a given sleep stage. We called this metric **Time Deviation**, which, for an algorithm Alg, measures the average difference on the number of epochs assessed as in sleep stage s ($TimeInSleepStage_s$) between Alg and the ground truth (GT). This can be defined, for a study with Q participants as:

$$TD_s(Alg, GT) = \frac{1}{Q}\sum_{i}^{Q} TimeInSleepStage_s(Alg, i) - TimeInSleepStage_s(GT, i)$$

Many sleep classification studies have reported accuracy and F_1 for the proposed models, where these indicators measure the predictive performance of the algorithm based on a time indexing manner without directly summarizing the total time at each sleep stage throughout the night. However, for clinicians and other health practitioners, confusion matrices are not the most obvious way to present the time deviation of sleep stages that predicted by each classifier. Here we introduce time deviation as a complementary metric to what is offered by confusion matrices, allowing healthcare practitioners to gain an intuitive understanding of how some stages are either over or under estimated. The time deviation metric calculates

the difference between the predicted minutes for each sleep stage and the ground-truth PSG annotation. Additionally, to understand individual sensor contributions within our multimodal set-up, we evaluated feature importance and the effect of features on the different models and classification tasks using SHAP analysis [360, 361]. SHAP values allow explaining the output of a function f as the sum of the effects ϕ_i of each individual feature that is introduced to a conditional expectation [360]. Shapley values can be defined as follows:

$$\phi_i(N) = \frac{1}{|N|!} \sum_R \left(v(P_i^R \cup \{i\}) - v(P_i^R) \right) \ ,$$

where ϕ is the Shapley value, N is the number of features, P_i^R is the set of features with order, $v(P_i^R)$ is the contribution of the set of features with order and finally $v(P_i^R \cup \{i\})$ is the contribution of the set of features with order and feature i. By using Shapley, SHAP returns individual contributions to the full feature set in each particular task we introduced allowing us to understand individual feature contributions not only for a particular task, but also for each sleep stage. Here, to understand global importance of the features imputed into the models, we average the absolute Shapley values per feature across the data and then sort those features by decreasing importance.

Similarly, to gain a better understanding of how each classifier performs at each task beyond conventional statistical learning metrics, we include a measure of the total deviation in time on a model-by-model basis for the class (stage) predicted and standard deviation. For instance, a classifier might be prone to over-classify NREM sleep and miss out on epochs that were supposed to be Wake or REM.

3.5 Results

For our experiments, we used a total of 1,743 nights of sleep, representing 1,903,900 sleep epochs of 30 seconds. The prevalence of sleep stages (AASM convention used) within these epochs is reported in Table 3.6.For consistency, all of our architectures and models were evaluated during sleep recording period across tasks, with performances reported in Table 3.7 for binary classification and Table 3.8 for multistage classification. Within each table, results were sorted by mean accuracy in descending order. In this section, we only show the top three DL classifiers alongside the best classifier from ML. Table 3.10 provides a summary of sleep measured by PSG and the time spent in different sleep stages for the best classifier in each task.

3.5.1 Task 1: sleep-wake classification

The best performing algorithms for Task 1 are presented in Table 3.7, and a full breakdown of all classifiers is presented in the supplementary tables for this task. We used baseline

Table 3.6 Number of 30-seconds sleep epochs for each of the four tasks studied in this work. The numbers in parentheses were obtained within sleep period time which measured from the first to the last non-wake detected sleep epoch.

Task 1			**Task 2**			**Task 3**			**Task 4**		
Sleep Stages	# Epochs	%	Sleep Stages	# Epochs	%	Sleep Stages	# Epochs	%	Sleep Stages	# Epochs	%
Wake	652,509(314,784)	34%(20%)	Wake	652,509(314,784)	34%(20%)	Wake	652,509(314,784)	34%(20%)	Wake	652,509(314,784)	34%(20%)
Sleep	1,251,391	66%(80%)	NREM	1,022,346	54%(65%)	Light	893,472	47%(57%)	N1	171,027	9%(11%)
			REM	229,045	12%(15%)	Deep	128,874	7%(8%)	N2	722,445	38%(46%)
						REM	229,045	12%(15%)	N3	128,874	7%(8%)
									REM	229,045	12%(15%)
Total	1,903,900(1,566,175)	100%	Total	1,903,900(1,566,175)	100%	Total	1,903,900(1,566,175)	100%	Total	1,903,900(1,566,175)	100%

Table 3.7 **Sleep wake classification results (mean $\pm$ standard error at 95% confidence interval) and predicted minutes by multimodal and single modality approaches (full recording period)**; Actigraphy modality: 🏃, HR/HRV modality: ❤; (*Full Table available on supplementary, **Average time deviation from ground truth across all subjects $\pm$ standard error)

Sleep-Wake Classification Benchmarks*									
Method Specifics			**Performance Metrics**						**Time Deviation****
Modality	Sensors	Top 3 Classifiers	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Sleep (mins)
Multimodality	❤ 🏃 *[HR/HRV, Actigraphy]*	CNN (100)	**84.4 ± 1.0**	**67.9 ± 2.0**	**84.8 ± 1.3**	92.4 ± 1.2	87.6 ± 1.1	**62.0 ± 2.0**	36.2 ± 7.3
		LSTM (100)	**84.4 ± 1.0**	67.4 ± 1.9	84.7 ± 1.2	92.5 ± 1.1	**87.8 ± 1.0**	61.6 ± 2.1	**36.0 ± 6.7**
		CNN (50)	84.3 ± 1.0	67.3 ± 2.0	84.5 ± 1.3	**92.7 ± 1.2**	87.6 ± 1.1	61.7 ± 2.1	39.0 ± 7.2
		Random Forest (300)	82.3 ± 1.0	65.7 ± 2.1	83.7 ± 1.3	90.6 ± 1.1	57.6 ± 2.1	57.1 ± 2.1	32.9 ± 7.3
Single Modality	❤ *[HR/HRV]*	LSTM (100)	**79.5 ± 1.2**	**62.2 ± 2.1**	**81.8 ± 1.4**	88.9 ± 1.3	**84.1 ± 1.1**	**51.5 ± 2.2**	35.3 ± 4.4
		CNN (100)	79.1 ± 1.2	57.0 ± 2.1	79.9 ± 1.5	**91.0 ± 1.4**	83.9 ± 1.3	49.8 ± 2.1	54.4 ± 4.7
		LSTM (50)	78.6 ± 1.2	61.1 ± 2.0	81.2 ± 1.4	88.2 ± 1.3	83.4 ± 1.2	49.5 ± 2.1	**34.5 ± 4.6**
		Random Forest (300)	70.3 ± 1.2	39.2 ± 2.3	73.6 ± 1.4	86.7 ± 1.9	77.6 ± 1.5	27.1 ± 1.7	70.4 ± 12.5
	🏃 *[Actigraphy]*	CNN (100)	**84.9 ± 1.0**	67.1 ± 2.0	84.7 ± 1.3	**93.8 ± 1.0**	**88.3 ± 1.0**	**63.0 ± 2.0**	43.0 ± 6.9
		CNN (50)	84.4 ± 1.0	67.6 ± 2.0	84.6 ± 1.3	92.9 ± 1.1	87.8 ± 1.1	62.2 ± 2.1	39.0 ± 7.1
		LSTM (100)	84.3 ± 1.0	**69.7 ± 1.8**	**85.5 ± 1.2**	91.2 ± 1.1	87.6 ± 1.0	62.0 ± 2.0	**26.5 ± 6.6**
		Random Forest (300)	81.2 ± 1.0	63.4 ± 2.0	82.9 ± 1.3	89.7 ± 1.1	85.4 ± 1.0	54.1 ± 2.1	32.6 ± 7.2

approaches, namely, *Always Sleep* and *Always Wake*, which showed that 66.5% of the epochs are sleep. Given the fact that for the purpose of this work, we only explored classification during the night period, we established that 66.5% was the minimum accuracy threshold for our models. Furthermore, although not reported on this work, we tested several of the well-established heuristic algorithms such as Cole-Kripke [82] and Sadeh [81] on our single-modality actigraphy data. Our results agree upon what was reported by Palotti et al. [237]. All of these approaches were outperformed by both the feature-based ML and DL models explored in our work.

All traditional ML modalities showed similar performance. Corroborating what had been shown in the related work that, when these algorithms are applied to actigraphy data, they result in high sensitivity but poor specificity [237]. Interestingly, for this task, adding HR and HRV to actigraphy for a combined sensing modality on the top classifier of CNN (100) did not significantly improve F_1 ($p = 0.347$), accuracy ($p = 0.499$) or Cohen's κ ($p = 0.506$). As expected, HR/HRV alone did not yield comparable performance to actigraphy alone or the combined sensing approach.

Table 3.8 **Sleep stage classification results (mean ± standard error at 95% confidence interval and predicted minutes by multimodal and single modality approaches (full recording period)**; Actigraphy modality: 🚶, HR/HRV modality: ❤; Three different tasks: Task 2: 3 stages, Task 3: 4 stages, Task 4: 5 Stages (*Full Table available on supplementary, **Average time deviation from ground truth across all subjects ± standard error)

Task 2: Wake, NREM, REM

Method Specifics			Performance Metrics						Time Deviation*		
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	NREM
Multimod.	❤ 🚶	LSTM (50)	**76.2 ± 1.0**	**85.6 ± 0.5**	**72.2 ± 1.3**	68.8 ± 1.2	67.9 ± 1.3	58.4 ± 1.8	-13.2 ± 6.8	-10.7 ± 3.8	**23.9 ± 7.1**
		LSTM (100)	76.1 ± 0.9	85.1 ± 0.5	71.9 ± 1.4	66.8 ± 1.2	66.4 ± 1.3	57.4 ± 1.9	**-3.2 ± 6.8**	-23.3 ± 3.4	26.5 ± 7.0
		CNN (100)	76.0 ± 1.0	**85.6 ± 0.6**	**72.2 ± 1.2**	**69.7 ± 1.3**	**68.1 ± 1.3**	**58.6 ± 1.9**	-32.7 ± 7.2	**2.5 ± 4.5**	30.2 ± 7.7
		Random Forest (300)	70.5 ± 0.9	79.9 ± 0.5	59.2 ± 1.5	53.0 ± 0.7	50.3 ± 0.7	47.6 ± 1.7	-20.0 ± 7.6	-63.6 ± 2.9	83.5± 7.8
Single Modality	❤	LSTM (100)	**73.8 ± 1.2**	**84.3 ± 0.6**	**69.8 ± 1.5**	**66.1 ± 1.3**	**64.9 ± 1.5**	**50.0 ± 2.2**	-27.8 ± 8.5	-8.6 ± 4.2	36.4 ± 8.1
		LSTM (50)	72.9 ± 1.1	83.8 ± 0.6	67.9 ± 1.4	64.1 ± 1.3	62.9 ± 1.4	45.5 ± 2.1	-17.9 ± 8.5	-16.2 ± 4.3	34.1 ± 8.3
		CNN (100)	71.0 ± 1.2	83.6 ± 0.6	66.3 ± 1.4	65.4 ± 1.4	62.7 ± 1.4	46.1 ± 2.0	**-12.9 ± 9.4**	**2.3 ± 5.0**	**10.6 ± 9.1**
		Random Forest (300)	59.6 ± 1.0	73.8 ± 0.4	48.4 ± 1.4	43.4 ± 0.6	39.2 ± 0.8	19.7 ± 1.4	-26.8 ± 13.5	-65.3 ± 3.1	92.0 ± 13.2
	🚶	LSTM (100)	**71.4 ± 0.9**	**80.1 ± 0.6**	**51.7 ± 1.1**	**52.9 ± 0.7**	**49.8 ± 0.8**	**49.7 ± 1.7**	**-16.8 ± 7.3**	**-67.0 ± 3.0**	**83.8 ± 7.7**
		CNN (100)	71.0 ± 1.0	79.5 ± 0.6	50.1 ± 0.8	52.1 ± 0.8	49.1 ± 0.8	48.0 ± 1.8	-34.9 ± 7.5	-67.6 ± 3.0	102.5 ± 7.9
		LSTM (50)	70.9 ± 0.9	79.7 ± 0.6	49.0 ± 0.8	52.4 ± 0.7	49.2 ± 0.8	48.3 ± 1.7	-20.3 ± 7.4	-67.6 ± 3.0	87.9 ± 7.8
		Random Forest (300)	68.6 ± 0.9	78.8 ± 0.5	53.4 ± 1.1	51.0 ± 0.7	48.5 ± 0.8	44.3 ± 1.7	-20.5 ± 7.2	-60.7 ± 2.9	81.2 ± 7.4

Task 3: Wake, Light Sleep, Deep Sleep, REM

Method Specifics			Performance Metrics						Time Deviation*			
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	Deep Sleep	Light Sleep
Multimod.	❤ 🚶	LSTM (50)	**70.3 ± 1.0**	**87.4 ± 0.4**	57.9 ± 1.3	**54.0 ± 1.0**	**51.9 ± 1.0**	**53.8 ± 1.9**	**-1.0 ± 6.9**	-5.6 ± 4.0	-36.2 ± 3.5	**42.8 ± 7.4**
		LSTM (100)	70.2 ± 1.0	86.9 ± 0.4	**59.9 ± 1.5**	52.4 ± 1.0	51.3 ± 1.1	51.7 ± 1.8	-18.9 ± 6.6	-24.7 ± 3.7	**-32.4 ± 3.5**	76.0 ± 7.3
		CNN (100)	69.0 ± 1.0	87.0 ± 0.4	58.0 ± 1.4	53.7 ± 1.0	51.2 ± 1.1	51.6 ± 1.8	-15.9 ± 7.5	**4.4 ± 4.8**	-34.5 ± 3.5	46.1 ± 8.1
		Random Forest (300)	63.6 ± 1.0	83.3 ± 0.4	44.7 ± 1.3	40.1 ± 0.6	36.7 ± 0.6	34.4 ± 1.3	-15.2 ± 7.6	-61.3 ± 2.9	-38.9 ± 3.6	115.3 ± 8.3
Single Modality	❤	LSTM (100)	**67.4 ± 1.2**	**86.2 ± 0.4**	**56.2 ± 1.6**	**51.3 ± 1.1**	**49.5 ± 1.2**	**44.6 ± 2.2**	**-13.1 ± 8.4**	-13.5 ± 3.8	**-33.7 ± 3.5**	60.4 ± 8.1
		LSTM (50)	66.2 ± 1.1	85.6 ± 0.4	54.4 ± 1.5	49.5 ± 1.1	47.4 ± 1.1	41.2 ± 2.1	-15.0 ± 8.1	-14.6 ± 4.1	-36.4 ± 3.5	65.9 ± 7.9
		CNN (100)	64.3 ± 1.1	85.3 ± 0.4	54.4 ± 1.6	50.2 ± 1.1	47.1 ± 1.1	40.9 ± 2.1	-23.0 ± 9.5	**8.2 ± 5.1**	-34.8 ± 3.5	**49.6 ± 8.9**
		Random Forest (300)	53.3 ± 1.0	79.2 ± 0.4	35.5 ± 1.1	33.2 ± 0.5	28.6 ± 0.6	12.6 ± 1.1	-4.3 ± 14.0	-64.7 ±3.0	-39.0 ± 3.6	-39.0 ± 3.6
	🚶	LSTM (100)	**64.1 ± 1.0**	82.9 ± 0.5	35.6 ± 0.7	**39.6 ± 0.7**	**35.8 ± 0.7**	**33.5 ± 1.4**	-32.5 ± 7.4	-67.6 ± 3.0	**-39.3 ± 3.6**	139.4 ± 8.5
		CNN (100)	63.9 ± 1.0	**83.0 ± 0.4**	**36.3 ± 0.9**	**39.6 ± 0.7**	35.7 ± 0.7	**33.5 ± 1.4**	**-26.4 ± 7.6**	-67.5 ± 3.0	**-39.3 ± 3.6**	**133.2 ± 8.7**
		LSTM (50)	63.6 ± 1.0	82.7 ± 0.4	35.6 ± 0.8	39.3 ± 0.7	35.5 ± 0.7	33.0 ± 1.4	-36.3 ± 7.1	**-67.3 ± 3.0**	**-39.3 ± 3.6**	143.0 ± 8.2
		Random Forest (300)	61.4 ± 1.0	82.6 ± 0.4	39.6 ± 0.9	38.1 ± 0.5	35.0 ± 0.6	31.2 ± 1.3	-15.6 ± 7.3	-59.3 ± 2.9	-37.1 ± 3.6	112.1 ± 8.1

Task 4: Wake, REM, N1,N2,N3

Method Specifics			Performance Metrics						Time Deviation **				
Modality	Sensors	Top 3 classif.	Accuracy	Specificity	Precision	Recall	F_1	Cohen's κ	Wake	REM	N3 Sleep	N2 Sleep	N1 Sleep
Multimod.	❤ 🚶	LSTM (50)	**63.7 ± 1.0**	88.7 ± 0.3	47.1 ± 1.4	43.0 ± 0.8	39.9 ± 0.8	56.3 ± 1.8	22.2 ± 7.1	-12.9 ± 3.9	-35.2 ± 3.5	**71.9 ± 7.5**	-46.0 ± 3.0
		LSTM (100)	63.6 ± 1.0	88.7 ± 0.3	47.8 ± 1.3	43.3 ± 0.8	40.5 ± 0.9	**57.0 ± 1.8**	**-3.3 ± 6.8**	-15.9 ± 3.9	**-32.3 ± 3.5**	97.7 ± 7.5	-46.2 ± 3.0
		CNN (100)	63.1 ± 1.1	**88.8 ± 0.3**	**51.5 ± 1.4**	**44.7 ± 0.9**	**41.9 ± 0.9**	56.2 ± 1.8	-26.2 ± 7.1	**8.2 ± 5.0**	-34.3 ± 3.5	92.4 ± 8.0	**-40.2 ± 3.1**
		Random Forest (300)	56.9 ± 1.0	86.2 ± 0.3	36.4 ± 1.2	33.1 ± 0.5	28.8 ± 0.5	46.3 ± 1.6	18.6 ± 8.1	-54.9 ± 3.1	-38.7 ± 3.6	123.6 ± 8.4	-48.6 ± 3.2
Single Modality	❤	CNN (20)	**55.6 ± 1.1**	86.4 ± 0.3	40.4 ± 1.2	37.3 ± 0.8	33.6 ± 0.9	36.2 ± 1.8	-15.6 ± 10.1	**-3.9 ± 5.8**	-39.1 ± 3.6	103.0 ± 9.8	-44.4 ± 3.0
		CNN (100)	**55.6 ± 1.1**	**86.7 ± 0.3**	**44.9 ± 1.4**	**38.9 ± 0.9**	**35.9 ± 1.0**	**37.1 ± 1.8**	**1.1 ± 10.7**	-12.0 ± 4.8	**-29.5 ± 3.6**	**81.2 ± 9.4**	**-40.8 ± 3.1**
		CNN (50)	54.2 ± 1.1	86.0 ± 0.3	41.2 ± 1.3	35.6 ± 0.8	32.1 ± 1.0	32.3 ± 1.9	35.7 ± 12.1	-28.2 ± 5.1	-36.4 ± 3.5	69.5 ± 11.0	-40.6 ± 3.1
		Random Forest (300)	46.6 ± 1.0	83.1 ± 0.3	29.9 ± 1.0	27.1 ± 0.4	22.3 ± 0.5	17.6 ± 1.4	48.5 ± 14.3	-61.3 ±3.1	-38.9 ± 3.6	97.8 ± 13.7	-46.2 ± 3.2
	🚶	LSTM (50)	**56.9 ± 1.0**	**85.7 ± 0.4**	26.1 ± 0.8	32.2 ± 0.6	27.1 ± 0.7	46.9 ± 1.7	-12.9 ± 7.5	-67.6 ± 3.0	**-39.3 ± 3.6**	169.2 ± 8.5	**-49.4 ± 3.2**
		LSTM (100)	**56.9 ± 1.0**	**85.7 ± 0.4**	25.3 ± 0.7	**32.3 ± 0.6**	27.1 ± 0.7	**47.1 ± 1.7**	**-3.3 ± 7.5**	-67.6 ± 3.0	**-39.3 ± 3.6**	159.7 ± 8.7	**-49.4 ± 3.2**
		CNN (100)	56.8 ± 1.1	85.8 ± 0.3	**27.7 ± 0.9**	32.2 ± 0.5	**27.2 ± 0.6**	46.9 ± 1.7	9.6 ± 8.3	**-65.6 ± 3.0**	**-39.3 ± 3.6**	**144.7 ± 9.1**	**-49.4 ± 3.2**
		Random Forest (300)	54.4 ± 1.0	85.6 ± 0.3	31.7 ± 0.8	31.1 ± 0.4	27.2 ± 1.6	42.7 ± 1.6	16.2 ± 7.6	-56.0 ±2.8	-36.7 ± 3.5	120.8 ± 8.0	-44.3 ± 3.3

3.5.2 Task 2: wake, Non-REM sleep, REM sleep classification

Task 2 evaluated sleep stages from a low granularity perspective by aggregating the different partitions of NREM. As observed in the supplementary table for this task, although some ML models had reasonable performance, the DL approaches were superior. It is important to note that at this level of granularity, NREM is overestimated while REM is underestimated for almost all models except for CNN (50) and CNN (100). In contrast to what was observed in Task 1, all models explored have a higher specificity than sensitivity and accuracy higher than F_1 score due to the imbalanced dataset.

As reflected in the top part of Table 3.8, the best classifiers for this task were all DL models for all sensor modalities. These models were significantly better than the best traditional ML model (Random Forest), with $p < 0.001$ for all metrics evaluated. The best DL algorithm with respect to accuracy was LSTM (50) which was also statistically better than CNN (20) ($p < 0.001$) achieving an accuracy of 76.2%. However, it was not significantly better than CNN (100) in terms of F_1, sensitivity and specificity ($p = 0.364, p = 0.138, p = 0.063$ and

$p = 0.399$). Nevertheless, CNN (100) achieved the lowest mean time deviation with a 2.5-minute overestimation of REM sleep. Interestingly, for this task, most algorithms' specificity significantly improved (e.g. LSTM (50) $p < 0.001$, reaching a specificity of 86%) when compared to Task 1 with the exception of the perceptron model.

In this task, it becomes apparent that multimodality is required for better performance at multistage classification, with the single modality approaches being significantly ($p < 0.001$) outperformed in all performance metrics and yielding much larger time deviations.

3.5.3 Task 3: wake, light sleep, deep sleep and REM-sleep classification

Task 3 explored sleep staging at a higher level of granularity than Task 2, with class imbalances being perhaps more apparent, as shown in Table 3.6. Here, DL approaches continued to outperform all feature-based ML models except for the Random Forest, which was not significantly worse than the CNN (20). The full results are available in supplementary tables. The best performing model was LSTM (50), although it was closely followed by LSTM (100) and CNN (100). Multimodal approaches were significantly better ($p < 0.001$) across all metrics upon comparison of the best classifiers for each category explored, depicting the value of these combined sensing approaches for multistage classification. Across all sensing modalities and all algorithms, deep and REM sleep were underestimated, with the exception of CNN (100) in the multimodal setup. In contrast, light sleep was overestimated, with wake being slightly underestimated across all setups, due to the class imbalance, except for LSTM (50).

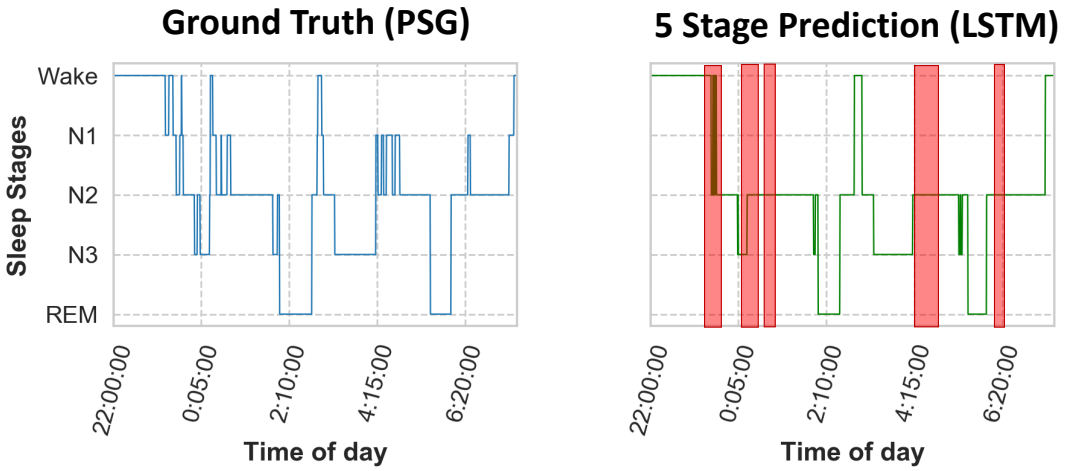
3.5.4 Task 4: wake, N1, N2, N3, REM sleep classification

Finally, Task 4 aimed to classify sleep stages following AASM rules (N1, N2, N3, REM and Wake). This task is the most complex due to its level of granularity and high-class imbalance, and as expected, the models performed worse here than in the previous tasks. An example of the best performing model LSTM (100) and the *mistakes* it makes is highlighted in Figure 3.4.

Like in previous tasks, the performance of DL algorithms was significantly better than feature-based ML algorithms as depicted in Table 3.8. The three best performing DL algorithms were not significantly different from each other with respect to accuracy($p > 0.05$) and F1 scores ($p > 0.05$). The best performing algorithm was LSTM (50), with an accuracy of 63.7% and an F_1 score of 39.9%. Even in the best performing multimodal approach, N2 tended to be severely overestimated (71 minutes more on average across the population). Nevertheless, the multimodal and HR/HRV approaches were good at classifying wake and REM, with only moderate deviations in time for those classes.

It is important to note that although the performance in terms of accuracy for the single modality approaches was comparable, each method struggled or had strengths at very different things.

Figure 3.4 Classification performance for multimodal, 5 stage classification using LSTM. On the top, the ground truth PSG, at the bottom, the predicted stages by the model. Highlighted in red are areas where the model does poorly.



For instance, HR/HRV was significantly better at classifying REM sleep in this modality than actigraphy. Similarly, upon evaluation of the algorithms only during the sleep period only (Table **??**, multimodal approaches were significantly better at detecting awakenings, yielding a more accurate wake after sleep onset (WASO) metric.

Figure 3.5 shows the confusion matrix for the best classifiers per task, allowing us to observe how models have an *easier* time classifying REM and wake and struggle to classify N1 and N3 (NREM). The observed time deviation in minutes substantiates this finding.

Finally, we evaluated the performance of different ensemble methods for each task. To validate the performance of our ensemble model, we conducted t-test based on both subject level as well as the group of subjects level. The difference between the two experiments lies in that for the second approach, we randomly divide all test subjects into 29 groups, each group containing 12 individuals. The purpose is to test whether the benefits of using ensemble methods are due to random chance.

The results of the two ensemble architecture models explored (based on different score-level fusion approaches) are shown in Table 3.9. We found no significant differences between the two ensemble models for all performance metrics assessed. However, they achieve better accuracy than single classifier approaches for all tasks and are significantly better on several performance metrics.

In Task 2, the ensemble approaches significantly outperformed LSTM (50) in terms of accuracy ($p < 0.05$), F_1 score ($p < 0.05$) and Cohen's κ ($p < 0.05$) and CNN (100) in terms of accuracy

Figure 3.5 Confusion matrix for the best classifier per Task

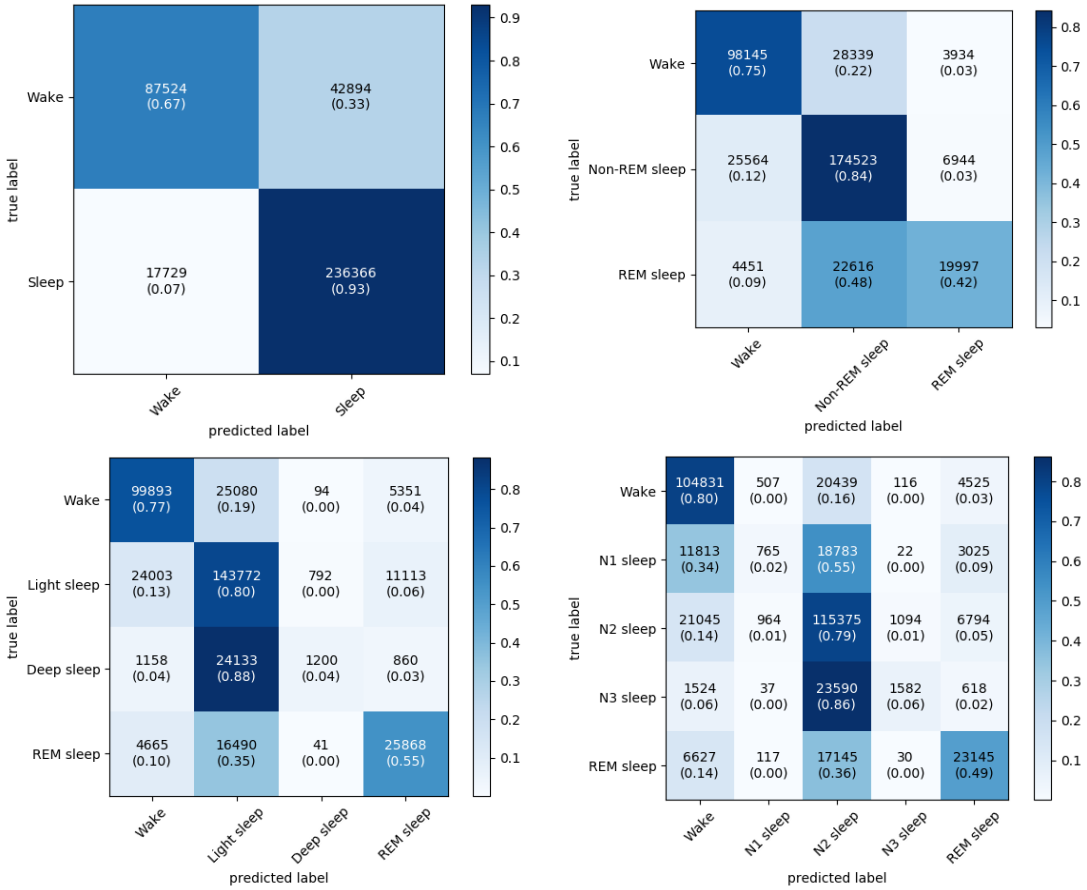


Table 3.9 **Results (mean ± standard error at 95% confidence interval) of different ensemble methods for each task.**(Mean over classifiers and Maximum selection are ensemble models)

Ensemble method	Accuracy	Cohen's κ	F_1	Precision	Recall	Specificity	Time Deviation (mins)			
Task 1 (2 Stages)							**Sleep**			
Maximum selection	85.3 ± 1.0	**64.4 ± 2.0**	88.4 ± 1.1	**85.7 ± 1.2**	92.8 ± 1.1	**70.1 ± 1.9**	**33.4 ± 6.7**			
Mean over classifiers	**85.4 ± 1.0**	64.3 ± 2.1	**88.5 ± 1.0**	85.4 ± 1.3	**93.4 ± 1.1**	69.1 ± 2.0	37.5 ± 6.8			
Task 2 (3 stages)							**Wake**	**REM sleep**	**NREM sleep**	
Maximum selection	77.9 ± 1.0	61.4 ± 1.8	69.6 ± 1.3	74.5 ± 1.3	70.6 ± 1.2	**86.5 ± 0.5**	**-16.1 ± 6.8**	**-11.1 ± 4.2**	**27.3 ± 7.5**	
Mean over classifiers	**78.2 ± 0.9**	**61.9 ± 1.8**	**69.8 ± 1.3**	**75.2 ± 1.3**	**70.7 ± 1.3**	**86.5 ± 0.5**	-21.7 ± 6.8	-13 ± 4.2	34.7 ± 7.5	
Task 3 (4 stages)							**Wake**	**REM sleep**	**Deep sleep**	**Light sleep**
Maximum selection	71.1 ± 1.0	55.7 ± 1.8	**52.4 ± 1.0**	**58.3 ± 1.3**	**54.8 ± 1.0**	**87.7 ± 0.4**	**-11.1 ± 6.7**	**-3.5 ± 4.6**	**-37.4 ± 3.5**	**52.0 ± 7.8**
Mean over classifiers	**71.6 ± 1.0**	**56.1 ± 1.8**	52.1 ± 1.0	57.1 ± 1.2	54.3 ± 1.0	87.6 ± 0.4	-17.8 ± 6.7	-8.9 ± 4.4	-38.5 ± 3.5	65.2 ± 7.8
Task 4 (5 stages)							**Wake**	**REM sleep**	**N3 sleep**	**N2 sleep**	**N1 sleep**
Maximum selection	65.2 ± 1.0	59.6 ± 1.8	**41.4 ± 0.8**	**49.7 ± 1.4**	**45.2 ± 0.8**	**89.3 ± 0.3**	**3.0 ± 6.9**	4.7 ± 5.1	**-37.1 ± 3.5**	**76.4 ± 7.8**	-47 ± 3.1
Mean over classifiers	**65.4 ± 1.0**	**60.1 ± 1.8**	41.2 ± 0.8	48.6 ± 1.4	44.9 ± 0.8	89.2 ± 0.3	-5.1 ± 6.9	**-2.1 ± 4.8**	-38.4 ± 3.5	91.9 ± 7.8	**-46.3 ± 3.1**

($p < 0.05$) and Cohen's κ ($p < 0.05$) based on both subject and group level t test. These models were the best two performers for that task prior to the introduction of the ensemble approach. Interestingly, on Task 4 (highest level of class granularity) the ensemble models only outperformed the best classifier (LSTM (50)) in terms of Cohen's κ and accuracy (for both subject and group level t test).

A summary of results per class and model are presented in Figure 3.6.

Figure 3.6 Performance (accuracy, F_1) per Task and model. Task 5 (ensemble architectures) are depicted against all benchmarks per each task on green

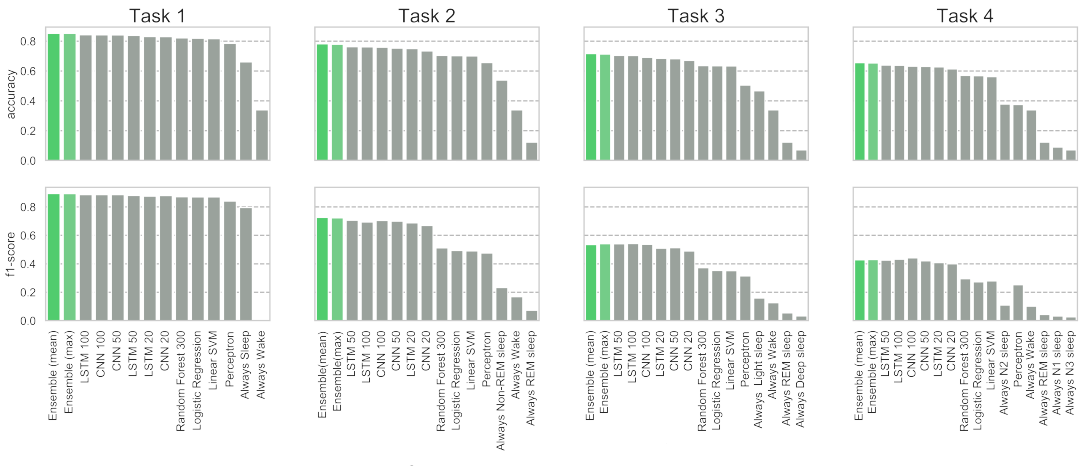


Table 3.10 **Sleep parameters and predicted minutes of each sleep stage in the *test* dataset.** Numbers are minutes except for the sleep efficiencies which are reported as percentages. Results are in mean +- SD/ and numbers in parentheses indicate the range in 95% CI (Mean over classifiers and Maximum selection are ensemble models)

Minutes of Sleep Stages						
Task	**Methods**	**Wake**	**Sleep**			
1	Ground truth	187.8±81.6 (179.2-196.4)	365.7±81.8 (357.1-374.3)			
	Mean over classifiers	150.2±73.2 (142.5-157.9)	403.3±92.0 (393.6-413.0)			
		Wake	**REM**	**NREM**		
2	Ground truth	187.8±81.6 (179.2-196.4)	68.6±27.2 (65.7-71.5)	299.7±66.5 (292.7-306.7)		
	Mean over classifiers	166.1±77.2 (158.0-174.2)	57.4±39.0 (53.3-61.5)	332.7±87.8 (323.5-341.9)		
		Wake	**REM**	**Deep Sleep**	**Light Sleep**	
3	Ground truth	187.8±81.6 (179.2-196.4)	68.6±27.2 (65.7-71.5)	41.8±33.7 (38.3-45.3)	258.8±65.7 (251.9-265.7)	
	Maximum selection	176.7±78.3 (168.5-184.9)	65.4±43.1 (60.9-69.9)	5.2±6.2 (4.5-5.9)	310.6±85.4 (301.6-319.6)	
		Wake	**REM**	**N1**	**N2**	**N3**
4	Ground truth	187.8±81.6 (179.2-196.4)	68.6±27.2 (65.7-71.5)	50.1±30.5 (46.9-53.3)	209.1±58.7 (202.9-215.3)	41.8±33.7 (38.3-45.3)
	CNN (100)	161.6±79.1 (153.3-169.9)	76.7±46.8 (71.8-81.6)	9.6±14.6 (8.1-11.1)	301.5±87.0 (292.4-310.6)	7.7±8.7 (6.8-8.6)

Sleep Parameters						
Task	**Methods**	**Total Sleep Time**	**Wake After Sleep Onset**	**Sleep Period Duration**	**Sleep Efficiency (Recording Period)**	**Sleep Efficiency (Sleep Period)**
	Ground truth	365.7±81.8 (357.1-374.3)	89.4±62.5 (82.8-96.0)	455.1±90.0 (445.6-464.6)	66.5±12.9 (65.1-67.9)	80.9±11.8 (79.7-82.1)
1	Mean over classifiers	360.7±84.0 (351.9-369.5)	70.9±57.4 (64.9-76.9)	474.2±92.9 (464.4-484.0	65.5±13.2 (64.1-66.9)	76.9±14.0 (75.4-78.4)
2	Mean over classifiers	359.5±84.3 (350.6-368.4)	81.7±60.1 (75.4-88.0)	469.1±93.0 (459.3-478.9)	65.3±13.2 (63.9-66.7)	77.4±13.9 (75.9-78.9)
3	Maximum selection	358.2±84.1 (349.4-367.0)	86.5±61.4 (80.0-93.0)	463.3±92.3 (453.6-473.0)	65.0±13.2 (63.6-66.4)	78.1±13.7 (76.7-79.5)
4	CNN (100)	361.1±83.3 (352.3-369.9)	94.6±68.6 (87.4-101.8)	486.5±90.7 (477.0-496.0)	65.6±13.1 (64.2-67.0)	75.0±14.2 (73.5-76.5)

3.5.5 Feature importance analysis

To understand how different modalities contribute to the prediction of each sleep stage, models that rank feature importance can be implemented. Many traditional ML approaches can provide feature importance ranking, such as logistic regression, linear Support Vector Machine or Random Forests. Of those, Random Forest is one of the most powerful traditional ML models, and it can rank feature importance by calculating the mean Gini impurity or mean

information gain over all its decision trees. However, these approaches only yield features that are important to the holistic classification task, and do not provide information on how these features contribute to recognizing certain classes (e.g., a sleep stage like REM sleep). We used SHAP [362] with Random Forest, which can generate class-wise feature importance. More technical details of SHAP can be found in [362].

By using this SHAP implementation with Random Forest, we can calculate the most important features per class, as shown in Figure 3.7. We report the top 20 features on Tasks 1-4, respectively. It is interesting to see how the top ranked features differ from task to task, pointing towards what contributes to more granular levels of classification.

For instance, in Task 1 (i.e., binary sleep/wake classification), the most informative features are from movement sensors (15 features out of 20), in contrast to those obtained from cardiac sensing (5 out of 20). However, cardiac features become more and more important as multi-stage classification tasks get more granular. In Tasks 2-4, with the increased class granularity, the most important features are cardiac features (8, 10, and 13 cardiac features, respectively) indicating the key role of cardiac sensing in distinguishing detailed sleep patterns.

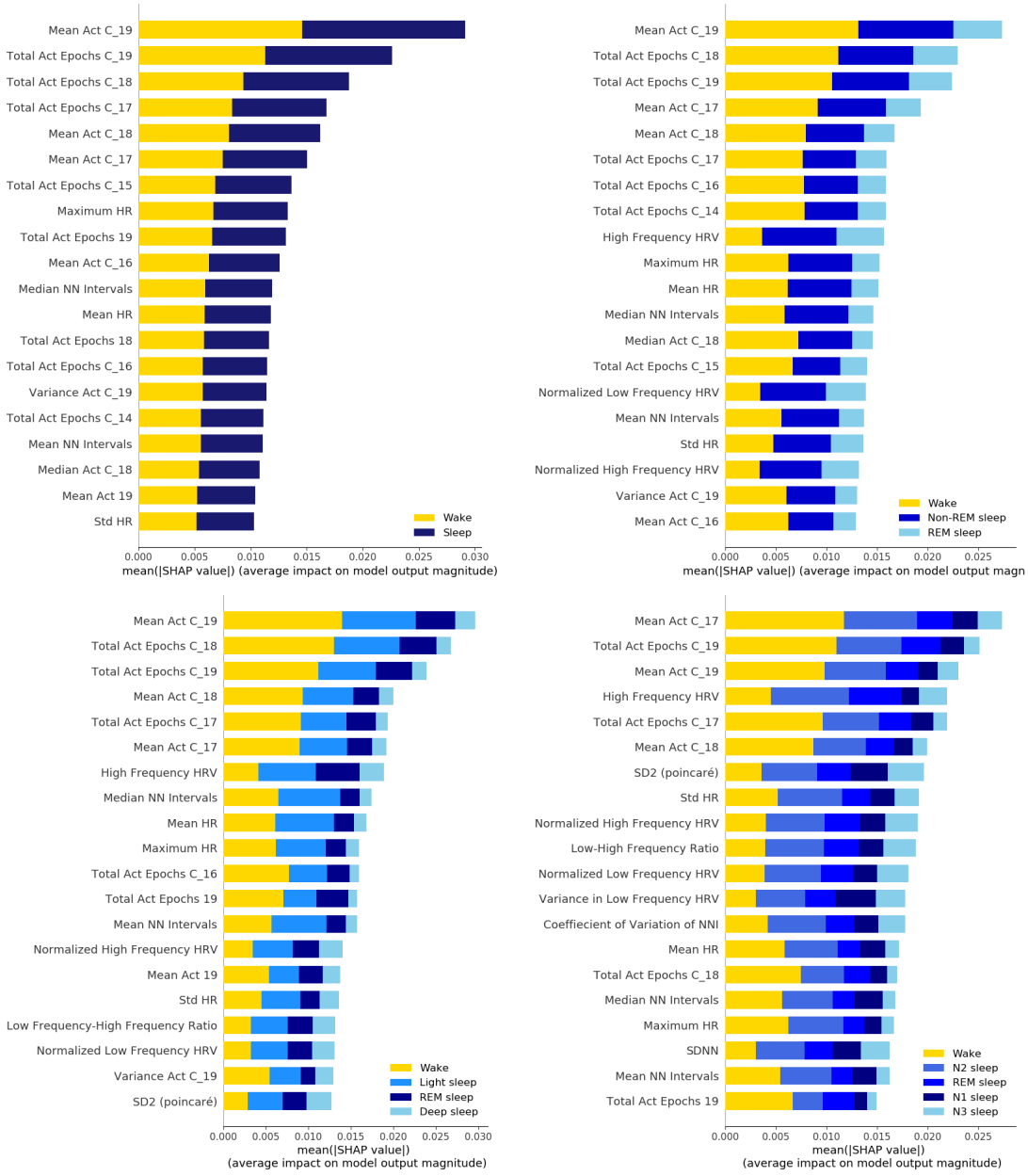
In multi-stage sleep classification (i.e., Tasks 2-4), it is also interesting to see feature importance associated to the different classes. Specifically, we observe high frequency HRV is the most discriminant feature in recognizing REM sleep, a finding that is consistent across all 3 multi-stage classification tasks. However, high frequency HRV is not as valuable in recognizing wake status. In Task 3 and 4, the non-linear HRV features such as SD2 which is the normalized Poincare plot parameter, SDNN, and coefficient of variation of NNI become more important than time domain features of HRV. Among these features, SD2 is ranked higher than many of the other HRV features except high frequency HRV in Task 4.

3.6 Discussion

3.6.1 Summary

This work presents the first systematic analysis of sleep-wake and sleep-stage classification using multimodal sensor data in a large, diverse population of both healthy and sleep-disordered participants. The main aim of this work was to understand how different models performed based upon details of the task (sleep-wake or multistage sleep classification) and sensor combination. To achieve this, we applied a series of traditional ML and DL approaches to each individual modality (i.e., actigraphy, ECG) and sensor combination (multimodal sensor fusion). Furthermore, we ran four different tasks to gain a deeper understanding of the strengths and limitations of the different approaches.

Figure 3.7 SHAP feature importance (Random Forest) for each Task is used to further understand feature contributions for each individual class (stage). The figure introduces the performance on Task 1-4, starting with binary sleep wake classification all the way to AASM's 5 stage classification



These tasks include: sleep-wake (Task 1); wake, NREM and REM (Task 2); wake, light sleep, deep sleep and REM (Task 3); and wake, N1, N2, N3 and REM (Task 4). The framework and analysis we provide were based on sensor modalities and signals that can be obtained from research grade wearable devices. Hence, RR-based metrics were used instead of raw ECG signals. Unlike raw ECG, these metrics may be derived from commercial research-grade wearable devices and in the near future also from non-clinical smartwatches that use both actigraphy/accelerometers and photoplethysmogram (PPG) [363]. With this work, we aim to

provide with a set of benchmarks for commercial and research studies and to inspire others to create open-access large population repositories to study the role of sleep and other physical behaviors in health and disease.

Here, we systematically evaluated how sensor modality affects classification outcomes and how model choice leads to differences in performance. Yuda et al. also explored a multimodal approach for sleep classification. Although their work is strong methodologically, the cohort is much smaller than that presented here, almost 70% of their cohort is male and the majority of their subjects had sleep disorders, limiting the generalisability of their findings [331]. Furthermore, they only explored the classification of three sleep stages. Here, we show that although multimodal sensor approaches do not lead to great improvements in classification performance for sleep-wake classification tasks, they are essential to classify sleep stages. For instance, actigraphy by itself struggles to classify REM sleep in Tasks 2-4 (Table 3.8), but it's performance improves when combined with HR/HRV. To-date, conventional sleep-wake classification algorithms have mostly exploited count-based movement data [81, 324] and models that combined HR information have mostly been confined to commercial devices based on device-specific algorithms.

Furthermore, this work highlights the strengths and limitations of the different models used, for instance, while CNN models do well at classifying high frequency transitions, LSTMs excel at classifying smooth patterns. LSTMs outperformed all other classifiers at multistage classification, due to their deep temporal modelling characteristics which align well with the multi-class, time-series classification problem that sleep-stages introduce. Meanwhile, CNNs were the best performers for binary sleep-wake classification tasks due to their ability to track exponentially longer sequences, such as those used in this type of task where the objective is less granular and has lower transition frequencies than the multi-class scenario. As such, our ensemble approach aimed to exploit individual model *strengths* to achieve better performance.

Remark: In sum, this work presents the first systematic analysis of single modality (actigraphy, HR/HRV) and multimodal sensing approaches for sleep-wake and sleep-stage classification using the most common feature-based ML and DL frameworks. Furthermore, a new ensemble architecture is introduced, outperforming all other models.

3.6.2 Transparency in algorithm development in machine learning for sleep health

All our analyses were performed in MESA [333, 334], a publicly available dataset for which access can be requested through: https://sleepdata.org/datasets/mesa. This creates transparency and facilitates reproducibility in human sleep science. We encourage others to use this resource to develop novel, more accurate models that exploit multimodal data. Here we provide an example of how the performance of well-established methods can be surpassed by an *ensemble*

architecture of DL models of different window sizes. Similarly, we found that the performance of our models was influenced by optimal hyperparameter search. In particular, we observed that although there was no significant improvement in terms of accuracy and F1 on our search space, certain patterns did emerge. For most CNNs and LSTMs, increasing the length of the sliding window improved performance (except for Task 1, sleep-wake only). Different search spaces may be able to yield more significant results and shall be explored in future work.

In this work, we advocate for and demonstrate the value of including performance metrics beyond the conventional accuracy, specificity, precision, recall and F_1 scores. For instance, introducing time deviation metrics allowed us to understand what precisely each model over- or underestimated. This is of particular value for the translational applications of this work which may be implemented by HCI researchers, clinicians or epidemiologists and have an impact on the field of digital health. These metrics are more interpretable for non-machine learning experts who may seek to understand how certain inferences should be interpreted. Clear, interpretable measures that allow non-specialists to understand the limitations of our models is critical both to the development of better study cohorts and to understanding the inferences made by these models.

3.6.3 Sleep classification performance by task

Binary sleep-wake classification (Task 1) using actigraphy had been previously explored by Palotti et al. in the same cohort [237]. Our results corroborate this study, with the CNN architecture narrowly outperforming the LSTM architecture. Interestingly, our multimodal approach did not add much to this binary classification task when exploring conventional metrics, but did yield a lower time deviation of overall sleep time than actigraphy alone. Models based only on cardiac signals had a slightly worse performance than both actigraphy alone and the multimodal approach, with accuracy estimates in the high 70s (79% for the LSTM (100)).

Task 2 consists of Wake, NREM and REM classification. This represents a valuable yet holistic overview of sleep stages and is what most free-living commercial devices aim to measure. Actigraphy and HR/HRV yielded an accuracy of around 74% and F_1 scores between 49% and 65%. Our time deviation metrics allowed us to observe that actigraphy cannot accurately determine REM sleep, whilst HR/HRV perform better. Both sensor modalities tend to overestimate time spent in NREM sleep. This finding also pertained to the multimodal approach, where NREM overestimation was the most error-prone estimate of the three states classified, at around 24 minutes mean deviation from gold-standard measures per participant on average when using the LSTM (100) and 24 minutes deviation when using the LSTM (50). NREM could be overestimated because it is the most common state among all participants on average, meaning that errors could be magnified. Accuracy estimates for our multimodal approach were in the high 70s for the majority of the classifiers, with LSTM (50) reaching

76% accuracy. These results are in line with the best performance previously reported in the literature. However, none of these previous studies had the scale and diversity that the MESA dataset offers [349].

Task 3 explored classification into Wake, REM, light sleep and deep sleep. In this classification, N1 and N2 were considered part of light sleep and N3 was classified as deep sleep. Actigraphy and HR/HRV reached accuracies of around 67% through LSTM (100). However, F_1 scores for the single modality approaches were between 35-50%. Similarly, both approaches overestimated time spent in light sleep and also struggled to pick up REM sleep. The multimodal approach outperformed the single modality approach with a higher accuracy of around 70% and an F_1 score of 52% for LSTM (50) (the highest performing model). Interestingly, LSTM (50) has a very strong performance at classifying wake and REM but struggles to discern light sleep and deep sleep, overestimating light sleep, probably given the transitional characteristics and high prevalence of the N2 stage nested in the light sleep class.

Task 4 aimed to evaluate classifiers that followed AASM scoring rules of Wake, REM, N1, N2 and N3. This task is the most complex of the four, given the high level of granularity required and the imbalance severity between sleep stages increased. Actigraphy and HR/HRV performances at this task were poor, with F_1 scores ranging from 27-36%. Both heavily overestimated the most prevalent state, N2. The multimodal approach struggled to discern among the different NREM stages and, again, overestimated time spent in N2. Its performance on Wake and REM was much better but, intriguingly, worse than what had been observed in Task 3. Accuracy did not exceed 64% and F_1 scores were between 40-42%. A visual illustration of this task is presented in Figure 3.4, where overestimation of N2 can be observed, alongside how the model struggles to discern transitions between N1, N2 and N3.

Across all multistage classification tasks LSTMs, outperformed every other modeling approach. This is most likely due to its ability to learn temporal dependencies from longer window sizes, contrasting with CNN models which focus on local dependencies. This makes LSTMs a particularly attractive candidate for multistage sleep classification given the intrinsic transitional nature of the task. The ensemble model approach serves as an example of how multistage classification benchmarks ought to be improved by new model architectures. The example approach is a rather simple one and thus only improves the performance marginally. Nevertheless, the results are promising. By incorporating different temporal domains and classifier types, these new models are able to pick up *nuances* that may have been tougher to identify by using a single convolutional or recurrent neural network.

Understanding sleep stage dynamics at a population level could be of value for digital health, epidemiology and clinical studies. We envision that research on behavioral change (i.e., [330]) can take advantage of ubiquitous systems for sleep stage classification to recommend changes for better sleep hygiene and healthier sleep architectures.

3.6.4 Physiological underpinnings of classifiers and sensor modality contributions

Our classification tasks aimed to explore how the different modalities performed with regards to the level of granularity and detail generated. Physiologically, sleep stages are quite different and one objective of this work was to understand the individual contributions of each sensor, as well as model biases and preferences. Following the AASM staging convention:

1. N1 (the first stage of NREM sleep) is the stage in which the change between wakefulness and sleep occurs. During this stage, heart rate, breathing and eye movement slow, with occasional muscle twitches. Similarly, slow-wave activity starts to appear on the PSG's EEG signal.

2. N2 (the second stage of NREM sleep) is the transition period between light and deep sleep. Heartbeat and breathing slow, muscles relax even further and body temperature drops. This stage is the most repeated across all sleep cycles. Together with N1, it is often referred to as *light sleep*.

3. N3 (the third stage of NREM sleep) is often referred to as *deep sleep*. Heart rate and breathing are at their lowest, muscles are very relaxed and it is rare for the person to awaken during this stage. These changes are also observed on the PSG's EEG signal, where the lowest frequency and highest amplitude waves can be found. Together with N1 and N2, this stage constitutes NREM sleep.

4. Finally, as explored in the introduction, REM sleep occurs in a cyclical fashion, approximately every 90 minutes. Breathing is faster and irregular, heart rate increases and, in healthy people, the body is in a state of temporary paralysis that prevents sharp movements related to dreams.

Given the physiological differences between sleep stages, we hypothesised that depending on the sensors used and the measurements they enabled, performance at classifying certain stages may differ. This is of high importance when considering the deployment of these technologies in clinical settings or for the exploration of the association between sleep characteristics and disease end-points in population-based research. Understanding time spent at different stages over a long-term period is of great importance for the greater sleep scientific community. For instance, during non-REM sleep, slow-wave activity has been shown to support memory consolidation [319] and reduce next-day anxiety [364]. In [319], for example, these slow-wave oscillations have been shown to affect the way the brain cerebrospinal fluid dynamics work, leading to oscillations in blood volume that draw this fluid across the blood-brain barrier.

We used SHAP to further understand how different sensor features contribute to the individual classification of sleep stages. Through this method, it became apparent that whilst actigraphy features were the most informative for sleep-wake classification, when moving to multistage

classification tasks, HR/HRV features were also important. This is reflected in Figure 3.7 where the top features contribute more to the model than the bottom ones, indicating their higher predictive power. For example, frequency domain features were very informative for recognizing non-REM sleep. Similarly, the application of this method to the different tasks allows for the direct comparison of feature importance across different levels of sleep architecture granularity. When exploring SHAP results, for Task 1, we found that the most informative features came from actigraphy, although maximum HR and NN intervals were also notable contributors. Activity coming from the wrist actigraphy was particularly important for wake classification. This finding carried through all 4 tasks and makes sense given the considerably higher amount of movement present during wake than in any sleep stage unless a sleep disorder is present. For Tasks 2–4, SHAP results helped us understand why the multimodal approach performs significantly better ($p < 0.001$) than individual sensors at sleep-stage classification. We observed that while HRV features were not particularly informative for wake classification, they added a lot of value to NREM and REM predictions. Interestingly, we observed that frequency domain HRV features were amongst the most informative for light and REM sleep classification, confirming our initial hypothesis derived from previous clinical reports [337, 338]. These findings emphasise the importance of including HR/HRV measurements combined with movement for multistage classification tasks using wearable devices.

3.6.5 Future work and limitations

The strengths of this work derive from its novelty, population size and generalisability. It is the first systematic assessment of multistage sleep classification using non-obtrusive sensors. This makes an important contribution to the literature with potential applications for clinicians, researchers and the wellness industry. The population used is uniquely diverse, including a breadth of racial backgrounds, balanced sex and a representative sample of sleep-disordered participants. This enhances the generalization capabilities of the findings in contrast to previous studies [341, 331].

Here, we only explored benchmark models, but more complex networks can be designed, for example, by training deeper architectures, bi-directional architectures or using attention mechanisms. A natural next step for this line of research would be to incorporate multi-task learning approaches as well as models that combine a convolutional base with LSTMs/bi-directional LSTMs on deeper layers. Similarly, given the temporal dependencies of these tasks, attention mechanisms may be well suited to improve model performance [365]. Another architecture that may yield interesting results is the addition of a dense layer to merge representations learned by RNNs and CNNs, also exploiting the unique contributions of each classifier (temporal representations by the RNNs and spatial representations by CNNs). Similarly, one potential avenue that is an alternative to the current approach would be to keep the same window size while using different model stride sizes.

Given the scope of our work and objectives, we did not explore in detail how different models perform in diseased versus healthy populations, or highlight the differences between them. Similarly, our models did not exploit the well-known reciprocal interaction model first introduced by McCarley and Hobson [366], which describes ultradian periodicity, the approximately 90-min sleep cycle, which indicates that NREM-REM stage transitions are regulated by both cholinergic and monoaminergic neuronal structures. This inherent sleep architecture shall be explored in future work to improve model performance. Furthermore, the ensemble model used in this chapter is a just an example of how our benchmarks can be improved by using a novel approach tool, many other models can be used inspired by what has been already done in automatic sleep scoring in EEG, as the tasks will follow the same temporal dependencies. Here, we did not explore bespoke deep network architectures neither covered the various possible approaches for multi-modal fusion. Our future work may consider exploring these areas of research.

Furthermore, given the scope of this work, we did not enforce strict quality control on the polysomnography data, which could have lead to the models to perform more poorly than if those practices and more stringent exclusion criteria had been applied. For instance, we found that a total of 30 subjects (about 2% of the total cohort) did not have any REM epochs at all. Similarly, on a small percentage of participants (less than 1%) accuracy scores were very low ($< 45\%$). After post-hoc visual inspection of those cases, we found that their sleep patterns were abnormal and five of them had a reduced number of sleep transitions. However, for the purposes of this work, those participant results were included in the final performance metrics. To exclude non-wear time and activity measurement failure from actigraphy data, we used human noted tags as the selection criteria. Full processing pipelines shall integrate automated non-wear time and data corruption detection algorithms in the preprocessing phase.

One limitation of this work is that the MESA cohort only includes adult participants. Thus the results cannot be generalised to teenagers or children. Further, as with all studies in this area to-date, all inferences are derived in laboratory settings, whilst the potential applications are in a free-living environment. Good quality "ground truth" data collection in free-living environments is complex and expensive although it is interesting to explore the possibility of larger studies using ambulatory PSG or wireless EEG for this purpose. Similarly, commercial, non-research grade devices have been shown to be unreliable at collecting longitudinal sleep measures [367]. Thus, we encourage these companies to be more transparent about the way they collect data and the algorithms they use.

3.7 Conclusion

In conclusion, this work introduces a systematic benchmark approach to sleep-wake and sleep stage classification using ML and DL approaches in single modal and multimodal settings. This approach advocates for model transparency, alongside reproducibility by exploring these

methods in the only open-access dataset, which includes diseased participants. The findings indicate that multimodal approaches combining movement with HR and HRV data were a valuable tool for the monitoring of sleep stages when those stages were aggregated to the level of NREM, REM, Wake. We further provide information regarding the performance of specific algorithms and guidance regarding algorithm selection depending on the classification tasks. Moreover, we introduce a deep ensemble model architecture which shows promising improvements in performance across the different multistage tasks explored. Overall, the findings highlight the promise of using wearable sensors as a low-burden, cheap and scalable approach for large, population-based studies. Future work should explore new model architectures to improve performance on more granular tasks, such as Tasks 3 and 4. These new architectures should aim to mimic the results obtained by EEG-based systems in both healthy and diseased populations as closely as possible. Furthermore, other auxiliary learning tasks and the inclusion of new, minimally obtrusive sensors may improve the performance of these models.

CHAPTER 4

DETECTING SLEEP IN FREE-LIVING CONDITIONS WITHOUT SLEEP-DIARIES: A DEVICE-AGNOSTIC, WEARABLE HEART RATE SENSING APPROACH

Publications

This work is under review for publication in Lancet Digital Health.

Contributions

I planned this project and devised the analysis plan in collaboration with my coauthors. In collaboration with Jõao and Marius I generated the code, conducted the statistical analyses and jointly interpreted the results. I wrote this chapter as well as the resulting manuscript incorporating feedback from all other co-authors.

4.1 Summary

Background: The rise of multisensor wearable devices offers a unique opportunity for the objective inference of sleep outside of laboratory environments, enabling scalable, longitudinal monitoring in large populations. To enhance objectivity and facilitate cross-cohort comparisons, sleep detection algorithms in free-living conditions should ideally rely on personalized but device-agnostic features, validated without the constraint of human sleep annotations or sleep diaries. Most commercial wearable devices can be used for sleep inferences but their algorithms are device specific, lack thorough validation against gold-standard measures and are not open-source. While on the previous chapter, we have shown that both Machine Learning and Deep Learning can yield strong results in supervise settings, many datasets and applications do not have these types of labels for use. In this Chapter we developed and validated a heart rate based algorithm that captures inter- and intra-individual differences in sleep, does not require annotations or diaries and can be applied in free-living conditions making it an optimal candidate for large cohort studies.

Methods: The algorithm was evaluated across four study cohorts, comprising over 2,000 nights of sleep using different research- and consumer-grade devices. The recording periods included both 24-hour free-living and conventional lab-based night-only data, facilitating the evaluation of our method under free-living and laboratory conditions. Our method was systematically optimized and validated against gold-standard polysomnography (PSG) and detailed sleep logs, and was compared to the results obtained from sleep period estimation arising from postural changes detected through accelerometry.

Results: We evaluated our approach in four separate cohorts comprising two free-living studies with detailed sleep logs and two PSG studies. In the free-living studies, the algorithm yielded a mean squared error (MSE) of 0.06 to 0.07 and a total sleep time deviation of -0.60 to -14.08 minutes. In the laboratory studies, the MSE ranged between 0.06 and 0.10 yielding a time deviation between -23.23 and -33.15.

Conclusions: Our results suggest that our heart rate-based algorithm can reliably and objectively infer sleep under longitudinal, free-living conditions, independent of the wearable device being used. This represents the first open-source algorithm to leverage heart rate data for inferring sleep without the need for sleep diaries or annotations.

4.2 Background

Human sleep is a physiological reversible state that is homeostatically regulated and vital for health and performance [368]. The functions of sleep are not fully understood but its influence on energy restoration, brain function, cognitive performance and behaviour, alongside interactions with the immune system, promotion of healing and consequences for numerous health conditions have been studied extensively [369–377]. As a consequence of its personal and public health significance, objective monitoring of sleep is paramount such that we can further understand its role in human health and behaviour. The gold-standard method to monitor sleep and diagnose most sleep disorders is PSG. PSG involves the collection and conveyance of different signals from many sensors operating simultaneously. Traditional PSG is limited to laboratory settings and requires an overnight stay for one or two days, expensive and obstructive equipment and trained laboratory technicians. These factors limit its scalability and prevent its use in objective sleep monitoring in large-scale population studies, as well as for long-term surveillance. Furthermore, the unfamiliar environment in which PSG monitoring takes place may result in atypical sleep that does reflect the study participants' or patients' typical pattern [378].

Chapter Significance: Sleep studies in free-living conditions are becoming easier to scale due to advances in sensor technology and affordability of consumer-grade wearables. Adoption has been driven by interest in long-term health monitoring, with incentives for person-generated health data insights. Historically, sleep classification algorithms have been trained on manually annotated records from selected cohorts with monophasic, night-time sleep data. Cultural preferences and shift work mean standard approaches might fail to reveal unbiased insights over 24-hour free-living conditions. In this chapter, we leverage the heart rate signal recorded by most state-of-the-art wearables and evaluated our device- and annotation-agnostic labeling approach across four separate study populations against gold- standard PSG or sleep diaries. Results show that heart rate-derived labels preserve the accuracy demanded in clinical sleep studies without the need for effort-intensive human annotations. This has potential applications in longitudinal studies and personalized medicine.

Actigraphy is a well-established and widely used method to objectively detect sleep non-obtrusively and longitudinally. This method, as well as its modern counterpart, accelerometry, are often integrated into wrist-worn wearable devices, offering a scalable and affordable alternative to PSG [379, 380]. Actigraphy has its precedent in early telemetric measurements of motor activity in the 1970s which were used to assess sleep quality [381]. Since then, a vast number of studies have assessed the use of actigraphy for sleep monitoring against PSG [379, 267, 382].

Over the past 30 years, a number of actigraphy-based algorithms have been derived with nocturnal sleep-wake scoring, showing strong validity and reliability against PSG [255, 383–387]. These algorithms have been readily used ever since and were recently benchmarked against both each other and newer machine learning and deep learning methods, highlighting the strengths and limitations of each method [223]. Across multiple studies that evaluated the performance of actigraphy against PSG, it was shown that actigraphy struggled to classify wake events during the sleep period, yielding poor specificity [223, 380, 386, 388, 389]. Similarly, these actigraphy-based algorithms were only optimised for nocturnal sleep-wake scoring, thus, they face a major challenge when applied to 24 hour recording. In order to work, they require additional information from expert sleep scorers or reliable sleep diaries [390, 112]. This requirement is also observed in most proprietary commercial brand algorithms, requiring the user to report habitual sleep times when they first set up a profile. Further, these algorithms do not allow the assessment of daytime sleep, severely limiting their relevance in cultures where multiple sleep episodes are common or amongst shift workers.

Wearable devices for both research and consumer applications have increasingly adopted multimodal sensing capabilities, combining movement and cardiac sensing, usually through actigraphy or accelerometry and photoplethysmography (PPG), respectively. These devices exploit recent advances in microelectromechanical systems (MEMS) and the associated improvements in cost, battery capacity, and increased memory. Thus, they are attractive not only for personal health monitoring, with large technological companies investing heavily in the space, but also in large epidemiological studies, as exemplified by the "All of US" research program [391]. Given the widespread adoption of consumer-grade wearable devices and their potential in large scale cohorts, the need for validation against gold-standard sleep measures has become imperative. Indeed, recent studies have shown that consumer and research-grade multimodal devices can be used to predict not only sleep-wake but also sleep stages during the night period [392, 393]. Furthermore, these multimodal approaches have been shown to improve the performance of models that only employ movement data in large populations, likely due to their ability to measure changes in autonomic nervous system activity reflected in heart rate (HR) and heart rate variability (HRV) [394]

Whilst these approaches are valuable, they have limited applicability in other large, free-living cohort studies for three main reasons. First, they rely on machine learning and deep learning methods which were derived specifically for those data sets, hence necessitating domain adaptation to be appropriately used in a different population or device. Second, in common with all previous well-established and widely used algorithms, they are only derived for the night period, limiting their ability to infer sleep in less regular sleepers, including shift workers [87]. Finally, these approaches rely on self-reported habitual sleep data, either through questionnaires (i.e., how much sleep you get on a typical weekday), or sleep diaries, which typically record times of getting into and out of bed. Both self-report measures are prone to recall bias, with survey data not providing sufficiently reliable labels [85, 395]. Of note, when using sleep diaries, it can often take in excess of 6 recorded days to achieve an agreement with

objective labels, even amongst those with more regular sleep patterns [86]. At present, the typical recording time for large-scale studies is about one week, contingent on device battery life. Thus, annotation and device/cohort-independent algorithms have the potential to infer sleep from objective 24-hour sensor data in a hitherto unprecedented manner.

In this work, we leverage heart rate data which can be obtained from most commercial and research grade wearable devices to develop a universal set of statistical attributes and an algorithm. We use these to infer both sleep periods and awakenings. In contrast to machine learning methods that require data to be sent to the cloud and large computational resources, our method is does not require this process or training before deployment, making it an attractive candidate to run directly on devices. This also limits the privacy issues associated with the transfer of personally generated data that are of paramount concern due to the nature of this data. The approach is also independent of device-type and make and was evaluated in four separate settings. First, the approach was developed in a large population (n=193) with multiple nights of recording accompanied by detailed sleep diaries. This cohort wore a combined heart rate and movement sensor, in addition to a variety of accelerometers (both wrist and hip), facilitating the comparison of the performance of our method against both the diaries and postural changes in multiple anatomical locations. We chose to develop the algorithm in this population because we were able to evaluate ≈ 8 nights of sleep per participant, enabling testing of both inter and intra-individual variability. We then assessed the performance of our method in a large, diverse, open-source dataset with PSG data (n=1,743). Moreover, to showcase the performance of our method in a readily available commercial device, we validated the approach in a cohort that wore an Apple Watch and concurrent PSG (n=31). Finally, the performance of our method in free-living conditions was further validated in a separate population against detailed, non-habitual sleep diaries that also wore a triaxial accelerometer and heart rate sensor (n=22).

4.3 Methods

Here we used four different datasets using a variety of devices and populations to showcase the performance of our proposed method.

4.3.0.1 The UK Biobank Validation Study (BBVS)

Participants of the BBVS study were recruited from the Fenland study [396]. In brief, 193 participants were recruited between the ages of 40 and 70, with a BMI between 20 and 50·kg · m^{-2}. Recruitment aimed to balance age, sex, and BMI distributions. Participants were invited to attend an assessment centre on two separate occasions, separated by a free-living period of 9 to 14 days during which they wore three waveform triaxial accelerometers (dominant and non-dominant wrists and thigh) as well as a combined movement and heart rate sensor. During the free-living period, participants were asked to keep a detailed log of their sleep, by recording the time they fell asleep and woke up on a daily basis. Ethical approval for the study was obtained from Cambridge University Human Biology Research Ethics Committee (Ref: HBREC/2015.16). All participants provided written informed consent. Full details of the BBVS study are described elsewhere [397].

4.3.0.2 Multi-Ethnic Study of Atherosclerosis (MESA)

The Multi-Ethnic Study of Atherosclerosis (MESA) is a multi-site prospective study that includes 6,814 men and women who identify as White, Black/African American, Hispanic, or Chinese, and are between the ages of 45-84 [333, 334]. Participants in this study were followed prospectively to evaluate risk factors for cardiovascular disease. 2,237 MESA participants are enrolled in a sleep exam (MESA Sleep Ancillary Study [398]), which includes seven days of wrist-worn actigraphy, one full overnight unattended polysomnography (wrist-worn actigraphy collected concurrently), and a sleep questionnaire. MESA participants who reported regular nighttime use of nocturnal oxygen or positive airway pressure devices were excluded from participation.

All data used from the MESA Sleep Ancillary study used in this work is publicly available from the National Sleep Research Resource repository[1]. Institutional Review Board approval was obtained at each MESA study site (Wake Forest University School of Medicine, Northwestern University, University of Minnesota, Columbia University, University of California Los Angeles and the Johns Hopkins University). All participants provided written informed consent.

[1] https://sleepdata.org/datasets/mesa

A number of common sleep disorders were identified and logged for the MESA sleep study, representing numbers that are close to their real prevalence in similar populations. A breakdown of those diseases is presented in Table 4.6.

4.3.0.3 PhysioNet Apple Watch Polysomnography Study

Data for this study was collected at the University of Michigan between 2017 and 2019. The study consisted of 39 healthy subjects with no prior diagnosis of sleep-related breathing disorders, parasomnias, restless leg syndrome, central disorders of hypersomnolence, peripheral vascular disease, cardiovascular disease, vision impairments not correctable by glasses or contact lenses or other disorders that could cause neurological or psychiatric impairment. The study also excluded on the basis of shift work and recent transmeridian travel. Furthermore, participants were ruled out on the basis of excessive daytime sleepiness according to the Epworth Sleepiness Scale, and after the PSG visit, participants which showed symptoms of either obstructive sleep apnoea or REM sleep behaviours were also excluded. A total of 31 subjects met the required criteria. Data for the study can be obtained through Physionet [399] and a detailed description of this data set is available elsewhere [392].

Participants in this study wore an Apple Watch to collect their activity patterns for 7 to 14 days before spending one night in a sleep lab. During the final night, participants underwent a PSG study while wearing the Apple Watch device (which collected HR and triaxial acceleration). The study was approved by the University of Michigan Review Board and all participants provided written informed consent.

4.3.0.4 The Multilevel Monitoring of Activity and Sleep in Healthy people (MMASH)

Data for the MMASH study was collected by BioBeats in collaboration with researchers from the University of Pisa and was obtained through Physionet [399, 400]. The study collected data from 22 healthy young adult male participants comprising continuous heart rate and triaxial accelerometry monitoring as well as variety of questionnaires to assess their physical activity, psychological and sleep characteristics as well as a detailed sleep diary. Participants also recorded their perceived mood (Positive and negative Affect Schedule-PANAS), Daily Stress Inventory (DSI) during the free-living protocol and completed a Morningness-Eveningness Questionnaire (MEQ), State-Trait Anxiety Inventory (STAI-Y), Pittsburgh Sleep Quality Questionnaire Index (PSQI) and Behavioural avoidance/inhibition (BIS/BAS) during their clinic visit. Further, anthropomorphic characteristics were recorded. All data was processed and recorded by sport and health scientists with the objective of assessing psychophysiological response to stress stimuli and sleep.

All participants provided written informed consent. Information was provided to them regarding the research protocol in accordance with General Data Protection Regulation: Regulation - EU 2016/679 of the European Parliament and of the Council 27/04/2016. Further, all experiments conducted were in accordance with the Helsinki Declaration as revised in 2013, the study was approved by the Ethical Committee of the University of Pisa (#0077455/2018).

Table 4.1 summarizes the types of wearable devices and ground truth used in each one of the studies.

4.3.1 Data processing

4.3.1.1 BBVS

Participants were fitted with a combined heart rate and movement sensor (Actiheart, CamNtech, Cambridgeshire, UK), measuring heart rate and uniaxial acceleration of the trunk every 15 seconds [401]. In addition, participants were fitted with three waterproof triaxial accelerometers (AX3, Axivity, Newcastle upon Tyne, UK); one device was attached to each wrist with a standard wristband, and one to the anterior midline of the right thigh using a medical-grade adhesive dressing. These devices were set up to record raw, triaxial acceleration at 100Hz with a dynamic range of $\pm 8g$. BBVS participants were asked to wear all four devices continuously for the following 8 days and nights while continuing with their usual activities. In addition, they were asked to complete a diary of their sleep onset and wake times daily. This ensured that any small changes in onset and offset of sleep were captured during the recording period.

Following the download of the devices, the combined sensor heart rate data was cleaned and non-wear periods identified by the combination of non-physiological heart rate and prolonged periods of no movement [402]. All signals from the triaxial accelerometers were re-sampled to a uniform 100Hz signal by linear interpolation, and then calibrated to local gravity using a well-established technique [403, 404]. Periods of non-wear were classified on the basis of windows comprising an hour or more wherein the device was inferred to be completely stationary, where stationary is defined as standard deviation in each axis not exceeding the approximate baseline noise of the device itself (13·milli-g). All non-wear periods were removed from the analysis. Additionally, pitch, roll and z-angles for all three accelerometry devices were calculated enabling postural assessments and direct comparisons to previously established approaches which only rely on acceleration data [113, 390]. The residual acceleration signal can be interpreted as a measurement of the rotated gravitational field vector which can then be used to determine the accelerometer's orientation angles (the conventional pitch and roll and z-angle, defined as the dorsal-ventral direction [113, 390]). Angles for each device were derived according to these formulae:

$$Pitch = \frac{tan^{-1}\left(\frac{Y}{\sqrt{X^2+Z^2}}\right) * 180}{\pi} \tag{4.1}$$

$$Roll = \frac{tan^{-1}\left(\frac{X}{\sqrt{Y^2+Z^2}}\right) * 180}{\pi} \tag{4.2}$$

$$Z\text{-}angle = \frac{tan^{-1}\left(\frac{Z}{\sqrt{X^2+Y^2}}\right) * 180}{\pi} \tag{4.3}$$

The accelerometry and heart rate signals were summarized to a common time resolution of one observation per 30 seconds and the time-series were aligned. Participants were excluded from the final analysis if they had less than 72 hours of concurrent wear data (three full days of recording from all four devices). Participants with less than 3 nights of concurrent wear and diary data were excluded from the final analysis. After these pre-processing steps the resulting analytical sample was of 158 participants. Three of these participants were on cardioreactive medication and two were taking betablockers.

4.3.1.2 MESA

The MESA Sleep Study was conducted using a Compumedics Somte System for PSG, which includes the ECG signals here used to derive HR and HRV and their associated features, alongside an Actiwatch Spectrum from Philips Respironics (Pennsylvania, USA) to record actigraphy data. This device captures measurements of movement defined as "activity counts"[2] and aggregates them into 30 second epochs. The Actiwatch was securely fastened to participant's non-dominant wrist. These actigraphy signals and their associated features can be derived in most research-grade wearable devices. The sensors for the Compumedics PSG comprised: cortical EEG, bilateral EOG, chin EMG, abdominal and thoracic respiratory inductance plethysmography, airflow, ECG, leg movement sensor and finger pulse oximetry. These sensors collected three types of signals: bioelectrical potentials (EEG, EOG, EMG, ECG), waveforms received from transducers (thermistors on the airflow devices, inductance respiratory bands, piezo leg sensors and position sensors from the leg device) and auxiliary devices (oximetry measures of oxyhemoglobin saturation and nasal pressure records). Full details of the setup, protocol and sampling rates are available[3,4]. All participants included in our study had at least one full night of PSG recording with concurrent actigraphy and ECG. All nocturnal recordings were transmitted to a centralized reading center at the Brigham and Women's Hospital (Boston, MA, USA) and data was scored by trained technicians using AASM guidelines.

For this study, we synchronized PSG, ECG and actigraphy records into 30-second sleep epochs for a subset of 1,743 out of the 2,237 participants included in the original study. A total of 494

[2] https://www.salusa.se/Filer/Produktinfo/Aktivitet/TheActiwatchUserManualV7.2.pdf
[3] https://sleepdata.org/datasets/mesa/pages/equipment/montage-and-sampling-rate-information.md
[4] https://sleepdata.org/datasets/mesa/files/documentation

participants were excluded on the basis of: (1) lack of concurrent PSG, ECG and actigraphy data; (2) lack of sufficient quality standard data ($<$ 3 h of usable data from the concurrent three sensing methods); or (3) lack of data integrity or misalignment of data, removing the resulting actigraphy outlier epochs based on human expert annotations. These outliers resulted from either non-wearing periods or equipment failure periods. For actigraphy epochs labeled as outliers, their corresponding HR/HRV epochs were also removed [405]. Further, to ensure that our evaluation was fair, we only included participants who had at least 30 minutes of wake time prior to sleep onset and a maximum of 240 minutes after sleep offset, resulting in a total of 1,183 participants.

To obtain HR information, we used the QRS complexes (R-points) detected using Compumedics Somte (Abbotsford, VIC, Australia) software Version 2.10 (Builds 99 to 101). The R-points were classified as normal sinus, supraventricular premature complex or ventricular premature complex. For the data cleaning, filtering and noise removal, we used the Python package HRV-analysis[5]. First, RR interval outlier data was filtered using a threshold method, with a range between 300 to 2000 ms, based on the approach previously described by Tanaka et al. [351], then ectopic beats were removed by through the methods described in Malik et al. [352]. After this step was completed, we linearly interpolated the removed R-points and we grouped the RR intervals into 30 seconds epochs.

4.3.1.3 PhysioNet Apple Watch

For the PhysioNet Apple Watch study, Apple Watch raw triaxial acceleration data (x,y,z axis measured in g) at a 50Hz resolution was converted into postural based metrics like the ones described on BBVS.

The Apple Watch measures HR in beats per minute, sampling every several seconds through its PPG sensor. For our analysis, we down-sampled HR to 15-second resolution. For the PhysioNet Apple Watch study, the laboratory technicians started a "recording" period for the watch before the PSG recording started. For our final analysis, we only included participants whose sleep onset and offset were greater than 10 minutes from the start and end of the recording period, respectively. Through this process we intended to introduce a more realistic setting for our model. Details on the laboratory PSG settings can be found elsewhere [392]. The final cohort consisted of 22 partipants.

4.3.1.4 MMASH

The 22 MMASH participants were fitted with two devices for continuous recording during 2 days: a heart rate monitor (Polar H7, Polar Electro Inc., Bethpage, NY, USA) which recorded

[5] https://pypi.org/project/hrv-analysis/

beat-to-beat intervals and was used to obtain HR data and a triaxial accelerometer (ActiGraph wGT3X-BT - ActiGraph LLC, Pensacola, FL, USA) was worn on the wrist. Participants were asked to wear the devices continuously during the duration of the protocol and to complete a diary of their sleep onset and wake up times during the recording period. For MMASH we followed the same pre-processing, data quality and noise removal protocols that we described in BBVS for both the triaxial accelerometry signal and the HR signal. Two participants were removed from analysis on the basis of missing diary entries.

A description of the cohorts we analysed and the wearable devices used to record data in each study is available in Table 4.1.

Table 4.1 **Summary of population size and devices used in the different datasets.**

Study	# Participants	Sensor type	Wearable device make	PSG	Sleep Diary
Biobank Validation Study	158	Triaxial accelerometer (3) Wearable ECG	AX3, Axivity (Newcastle,UK) Actiheart, CamNtech (Cambridge,UK)		✓
Multi-Ethnic Study of Atherosclerosis	1154	Actigraphy monitor ECG	Actiwatch Spectrum, Philips Respironics (PA,USA)	✓	✓
PhysioNet Apple Watch	22	Triaxial accelerometer Heart rate sensor (PPG)	Apple Watch (Series 2,3), Apple (CA, USA)	✓	
Multilevel Monitoring of Activity and Sleep in Healthy people	20	Triaxial accelerometer Heart rate sensor	ActiGraph wGT3X-BT, ActiGraph LLC (FL,USA) Polar H7, Polar Electro Inc (NY,USA)		✓

4.3.2 Algorithm to estimate the sleep window using heart rate

Several challenges must be accounted for when developing a method for the detection of sleep in free-living conditions. First and foremost, most methods derived for sleep-wake classification using wearable devices have been derived on and for use during the night period [385, 384, 223, 255, 392]. These approaches were mostly conducted in small studies using concurrent PSG and as such, their application during the full day period greatly compromises the quality of the results. They also tend to be optimized in small, non-diverse populations, comprising their generalizability to other cohorts. Moreover, they tend to be device and make specific, often requiring conversions into arbitrary activity intensity measures or counts. Finally, most algorithms that can be applied during the 24 hour period require sleep diaries or questionnaires for guidance, which are often biased and burdensome to obtain [406].

Here we introduce a simple approach to estimate sleep window leveraging the HR sensing capabilities that most modern wearables have. One of the major challenges presented by large cohort studies is inter-individual differences. For instance, individuals who are fitter, tend to have lower resting heart rates than those who are not as fit [99]. Hence, an approach that relies on HR signals should not follow a *one size fits all*, but rather adapt to each individuals' own heart rate profiles. To account for these considerations, we use the empirical cumulative distribution function (ECDF) of each individual's daily heart rate profile. This function, $F(x)$, is the probability that for each individual their heart rate takes a value x such that:

$$F(x) = P(X_i \leq x_0), \tag{4.4}$$

for every sequence $i = 1, ..., n$. Namely, $F(x_o)$ is the probability of the event $\{X_i \leq x_0\}$. In this case, x_0 is a threshold heart rate value (in beats per minute). To estimate the probability of a given event, we turn to the ratio of such an event given an individual's daily sample of heart rates. This results in:

$$\hat{F}_n(x_0) = \frac{\text{number of } X_i \leq x_0}{\text{total number of observations}} = \frac{\sum_{i=1}^{n} I(X_i \leq x_0)}{n} = \frac{1}{n}\sum_{n}^{i=1} I(X_i \leq x_0) \tag{4.5}$$

as the estimator of $F(x_o)$, that is the ratio of HR less than x_0, where $I()$ is the *indicator function*.

Thus, for every x_0, we can use such quantity as an estimator, so the estimator of the cumulative distribution function, $F(x)$ is $\hat{F}(x)$, which is referred to as the *empirical cumulative distribution function*.

By using the HR cumulative distribution function for each participant and each day of recording, our method accounts for inter- and intra-individual variation. It can adjust to different levels of fitness which often result in different resting HR during sleep [99]. Further, an elevated resting heart rate (RHR) accompanied by a fever is a well-known response to infection [407], alcohol consumption [408], stress [409] and can even be used to monitor influenza-like illness [410], something that our approach would account for. The method contains no in-built assumption of absolute time for the sleep window, and can therefore be used in night shift-workers and non-monophasic sleepers (those whose have more than one principal sleep windows in a 24-hour period) where the circadian HR rhythm is shifted so that most of the lower HR values still occur during sleep independent of the absolute time window when their sleep takes place. An example of our method applied to a shift worker can be observed in Supplementary Figure 4.9.

The first step of our heart rate sleep algorithm involves pre-processing the time series by assigning binary wake/sleep labels whenever the participant's heart rate dips above/below a specific quantile threshold (Q). The threshold value is calculated from the ECDF over 24-hour windows arbitrarily starting at 15:00 each day. Figure 4.3 showcases this cutoff for the full BBVS population based on two intervals (full day and from 21:00 to 11:00, a conventional night).

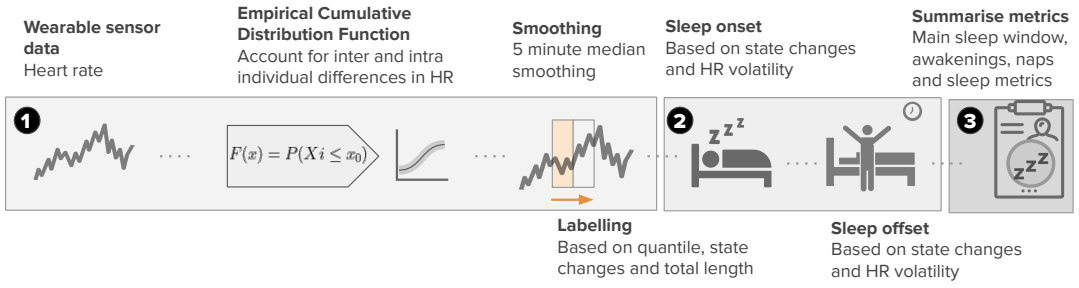
Wake/Sleep labels are then smoothed with a 5-minute rolling median and the length of their sequences is calculated. Sequences of sleep labels that are longer than a minimum length (L) are extracted and merged with other sleep sequences if their gap is smaller than a pre-defined

length (G). These Q, L and G parameters where optimized for each dataset with the goal of finding the best possible combination.

To be eligible as part of the final sleep window, the sleep sequence must not be preceded by more than 90 minutes of wake in the previous 4 hours of recording. The limits of the merged sleep sequences then guide a search (in a window starting 240 minutes before and 60 minutes after) for epochs with high HR volatility. This HR volatility threshold is defined as a rolling 10-epoch standard deviation of the HR signal of 6 beats per minute. Defining the final sleep window limits as the last, and first high volatility epochs for sleep onset and offset, respectively, is meant to increase the algorithm's sensitivity at discriminating sedentary time just before or after sleep (e.g. reading in bed) from the sleep window itself.

Finally, the algorithm also labels naps and awakenings, but these were not used in the analysis of the present datasets. Naps are the initial sleep sequences that lie outside a buffer 180-minute window either side of the main sleep window. For awakenings, the algorithm labels all the epochs when the HR rises above a quantile threshold AV extracted from the daytime (8am - 10pm) HR ECDF. From these only the sequences longer than 5 minutes are kept and then the sequences separated by less than 5 minutes of sleep are merged and then labeled as the final awakenings.

Figure 4.1 Heart rate sleep algorithm description. The approach can be broken down into three distinct steps. The first step, involves obtaining the wearable sensor HR data, pre-processing that data and setting initial sleep blocks through ECDF quantile thresholds Q. Blocks longer than L minutes are kept and merged with other blocks if their gap is smaller than G minutes. We extract the limits of the resulting blocks as sleep candidate for sleep onset and offset. Next, rolling heart rate volatility is used to refine these candidate times by finding nearby periods where this volatility is high. Finally, nap and awakenings are labeled, the former coming from the candidate sleep blocks not included in the largest sleep window, while the latter are short periods (<60 minutes) within the sleep window when the heart rate exceeds the daytime threshold. A detailed description of this algorithm and parameters used can be found in the methods section..



Pseudocode for the approach is provided in the Supplementary of this Chapter. A visual overview of the algorithm is provided in Figure 4.1 and Figure 4.2 showcases the application of the algorithm to a random participant trace.

Figure 4.2 Heart rate sleep algorithm in action for a participant chosen at random. The first step involves setting initial sleep blocks through ECDF quantile thresholds (in this experiment, $Q = .35$). Blocks longer than $L = 40$ are kept and merged if the gap between blocks is smaller than $G = 60$ minutes. We extract the limits of the resulting blocks as candidate state changes. The bottom panel highlights the use of rolling heart rate volatility to refine these candidate times by finding nearby periods where this volatility is high. The resulting candidate times designate each day's main sleep window.

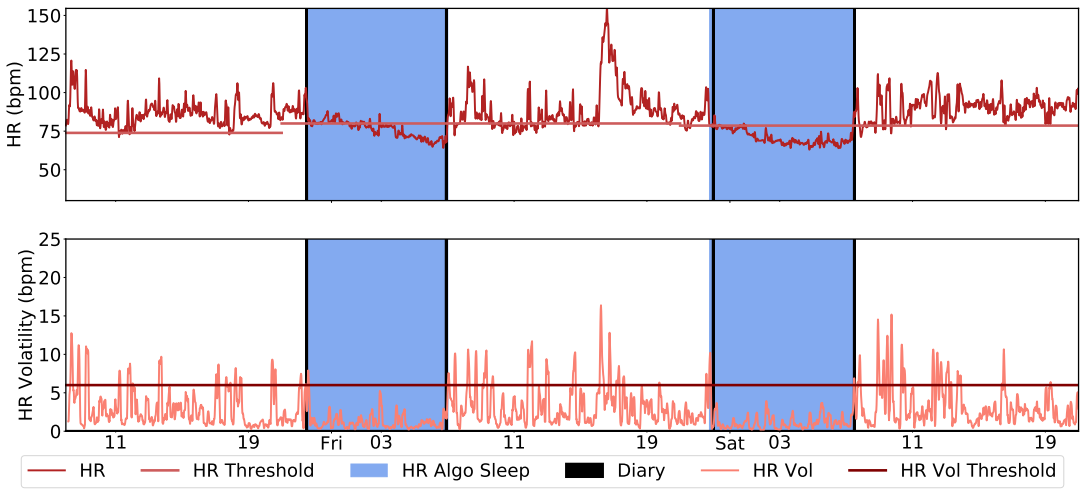
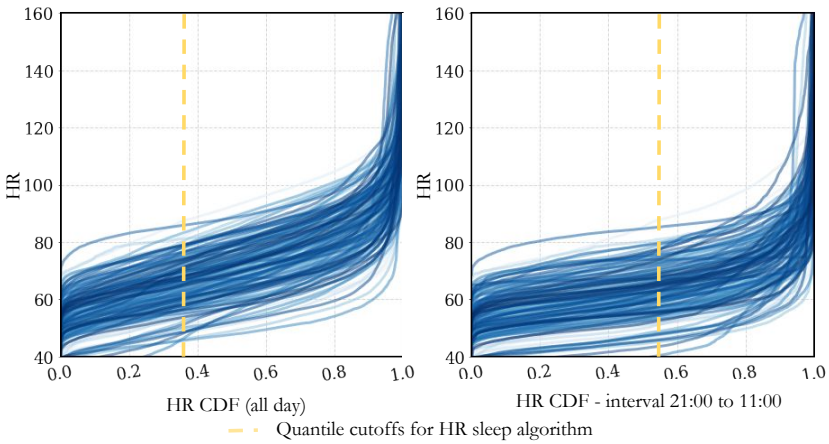


Figure 4.3 Cumulative distribution function for BBVS heart rates. (A) shows the HR ECDF for the full day across all participants and all days, (B) shows the HR ECDF for the periods of 21:00 to 11:00. The yellow dotted line shows the 0.35 cutoff for full-day and 0.55 for night-only and how for both time periods the HR profiles within those cutoffs are similar. Each individual line represents one participant for one day of recording.



4.3.3 Statistical analysis

4.3.4 Validation of the proposed approach

We used the four previously described cohorts to validate our method against gold-standard measures of sleep using PSG (MESA, Apple Watch PhysioNet) and detailed silver-standard

measures through sleep logs, as opposed to habitual sleep diaries which could be subject to recall bias (BBVS, MMASH). Although an ideal experimental protocol would have multiple days of PSG and free-living wearable sensor data, detailed sleep logs allowed us to evaluate the algorithm across more than one or two nights, showcasing the strength of our method at discerning both inter- and intra-individual variability.

We performed epoch by epoch evaluation on all four cohorts and derived comparisons regarding the performance of our method with regards to total sleep time (TST), sleep onset and sleep offset time.

4.3.4.1 Evaluation metrics

The following performance metrics were used to evaluate against the ground truth in each study: differences in onset, waking time, total sleep block duration (minutes), mean square error (MSE) and Cohen's kappa. We evaluated our algorithm systematically for individual HR CDF quantiles 0.25-0.9, window lengths of 25-50 minutes and time merge parameters between 0.5-6 hours, optimizing for MSE.

We defined MSE as:

$$MSE_{algo,diary} = \frac{\text{number of incorrectly labeled epochs}}{\text{number of epochs}} = \frac{\sum_{i=1}^{n}(algo_i - groundtruth_i)^2}{n}, \tag{4.6}$$

where $algo$ and $diary$ are the binary labels for that epoch (1 for sleep, 0 for wake) out of n minutes in each subject's heart rate time series. Epoch length is specified by the different study cohorts (1 minute in BBVS, 30 seconds in MESA and 15 seconds in both PhysioNet Apple Watch and 5 seconds in MMASH). Thus, if the sleep windows found by the HR algorithm match the ground truth labels exactly, $MSE = 0$. If the algorithm labels all epochs as wake, then MSE is the proportion of sleep in the time series according to ground truth, while if the algorithm and ground truth labels diverge entirely, MSE will be the sum of their sleep proportions out of the total time series. For all four cohorts we performed systematic parameter optimization for best MSE on the basis of quantile, window length and window merge values. We also computed cohen's kappa, which is used to determine the classifier agreement with ground truth (PSG or sleep diary), relative to chance [411]. Cohen's kappa is calculated through $(p_o - p_e)/(1 - p_e)$, where p_o stands for the percentage of observed classifications with agreement, and p_e is the percentage of classifications from hypothetical chance agreement. Finally, tests of statistical significance were conducted using a two-tailed t-test [412].

4.3.4.2 Evaluation with sleep diary and postural change: BBVS

In the BBVS study, participants wore a variety of wearable devices and recorded the time they went to bed and woke up on a daily basis, providing detailed sleep diaries. As such, we conducted two types of evaluations on this cohort.

Evaluation with sleep diary. First, we compared the performance of our method against those sleep diaries. For our evaluation, we only included participants who had filled out those diaries and had more than three days of concurrent sensing and diary data. We evaluated our model against the diaries in terms of total sleep time, sleep onset and offset.

Evaluation with angle change algorithm. We assessed the performance of our approach versus an angular change algorithm inspired by previous work [390, 113]. The angular change approach started with calculating the pitch and roll using triaxial acceleration for the device being evaluated. To isolate the gravitational acceleration for each axis, we applied a low-pass filter (0.2 Hertz) to each of the three axes (X, Y and Z) of every recording being evaluated.

Pitch, roll and z-angles were then calculated and the difference between successive epoch values was then smoothed using a 5 minute median rolling window. A threshold method ($<$ 10^{th} percentile of values in that given day $\cdot$ 15) was applied to both columns, dividing the time series into initial sleep and wake blocks.

Of these blocks, only those larger than 30 minutes were kept. Blocks separated by less than 60 minutes were then merged and the largest block was deemed as the main sleep block within the day.

Two different angular change evaluations were performed, first, the intersection of the epochs when both pitch and roll calculations agreed on a sleep label created a voting system for a more reliable final sleep window. Alternatively, z-angle only measures were used to generate those sleep metrics. All the previous steps were done separately for each limb (dominant and non-dominant wrists, and thigh) on which BBVS participants wore a device.

In BBVS, HR is recorded continuously across the 24-hr period. Thus, the threshold quantile is expected to be lower the longer the sampling interval for the ECDF given that sleep occupies a smaller proportion of the total interval being evaluated. To evaluate the effect of the chosen ECDF, we analyzed the optimal thresholds and parameters for both full-day and night-only (9PM to 11AM) and their associated results to better understand how experimental design may affect the performance of our approach.

4.3.4.3 Evaluation with polysomnography and sleep diary: MESA.

Evaluation with polysomnography. The recording time for PSG started when the subject's setup was complete, yielding a period of sedentary wakefulness prior to sleep onset. While in an ideal scenario the participant would have been subject to ground truth recording also during the day, this is not a possibility given the nature of PSG. However, this limitation was addressed by evaluating PSG against sleep diary on the same dataset and evaluating our approach against both PSG and diary data. For this evaluation we compared the resulting sleep blocks from PSG, defined as epochs where the participant was in either NREM (N1, N2, N3) or REM sleep, to the sleeping window obtained through our HR algorithm.

Further, in MESA, we explored how our algorithm performed in healthy participants versus participants with sleep disorders. To do so, we first evaluated in the full cohort (n=1,210) and then on the subset of participants with (n=199) and without (n=1011) any sleep disorders. The goal of this analysis was to caution and inform about potential limitations that our method may have when evaluating in diseased participants.

Evaluation with sleep diary. PSG derived sleeping windows were compared to sleep diary records in the MESA cohort. This comparison allowed us to further understand the deviations of habitual self-reported sleep to objectively monitored, ground-truth through PSG. For the evaluation we use the same metrics as previously explored in the evaluation against PSG.

4.3.4.4 Evaluation with polysomnography and postural change: PhysioNet Apple Watch Polysomnography Study.

Evaluation with polysomnography. The PhysioNet Apple Watch study provided a unique opportunity to test our method in a commercial-grade wrist-worn wearable sensor that was concurrently worn during PSG. For this study, we used the same evaluation method explored in MESA, exploring our method based on the night-time concurrent recordings of wearable HR and PSG.

Evaluation with angle change algorithm. Given the multimodal nature of the study, we evaluated both the HR based algorithm and the angular change based algorithm on this population. For this evaluation we followed the same procedure as previously described on BBVS.

4.3.4.5 Evaluation with sleep diary and postural change: MMASH.

In the MMASH study, participants wore an HR strap and triaxial wrist accelerometer and recorded detailed sleep diaries including the time they fell asleep and woke up, which was

filled on a daily basis. For this cohort, we also conducted two types of evaluation following the procedures used during the BBVS evaluation.

Evaluation with sleep diary. First, we compared the performance of our method against the sleep diaries of each participant. We evaluated our approach against the sleep diaries in terms of total sleep time, sleep onset and offset.

Evaluation with angle change algorithm. Similar to our second evaluation in BBVS, we also assessed the performance of our approach against the angular change approach previously described.

4.4 Results

Full descriptions of the population samples and devices included in our analyses are provided in Table 4.1.

4.4.1 Evaluation of the algorithm in the BBVS

The results of the evaluation on the BBVS study are summarized in Table 4.2. Our HR algorithm estimated TST on average 0.60 minutes less than those reported through sleep diary. For the full-day protocol the optimal quantile was 0.35 yielding an MSE of 0.06. On the other hand, the night-only HR analysis had an optimal quantile of 0.55 also resulting in an MSE of 0.06. The angular change approach had an underestimation of 125.37 minutes on the best performing wrist-worn device (non-dominant wrist). The results across all three accelerometers for this approach were comparable as summarized in Supplementary Table 4.7, each yielding an MSE of 0.10.

Our HR model estimated sleep onset on average 24.86 minutes later than sleep diary while the angular change approach resulted in an average of 85.4 minutes earlier. For sleep offset, our HR algorithm the estimation was on average 5.18 minutes earlier while for the angular change approach that estimation was 64.01 minutes for the device that yielded the best results (thigh). We also examined how night-only versus full-day evaluation influenced our results and resulted in different optimal parameters, particularly ECDF thresholds. These results are summarized in the Supplementary Table 4.8 and figures summarizing the results of this search for the BBVS cohort (both full-day and night-only) can be found in the Supplementary (Figures 4.7 and 4.8). Finally, modified Bland-Altman plots for the HR and angle approaches against sleep diary for the BBVS cohort are presented in Figure 4.4.

Table 4.2 **Comparison of HR and angle change algorithm performance for the BBVS dataset.** In this table, the angle change algorithm presented was applied on data from the device worn on the non-dominant wrist (ndw). Results for devices worn on other limbs are available in the Appendix.

Sleep parameter	**Metric**	**HR Algorithm** Value (mean ± 95% CI)	**Angle change algorithm (ndw)** Value (mean ± 95% CI)	**p-value**
Total sleep time	Time difference (minutes)	-0.60 ± 0.21	125.37 ± 0.26	< 0.00
	MSE	0.06 ± 0.00	0.10 ± 0.00	< 0.00
	Cohen's kappa	0.86 ± 0.00	0.76 ± 0.00	< 0.00
Sleep onset	Time difference (minutes)	1.14 ± 0.20	-60.17 ± 0.21	< 0.00
Sleep offset (Wake Up)	Time difference (minutes)	0.54 ± 0.16	65.20 ± 0.20	< 0.00

Figure 4.4 Modified Bland-Altman plot for BBVS. Modified Bland-Altman plot on the left shows the TST differences (delta) between the HR algorithm and diary in the y-axis and the x-axis shows the TST average for every participant. The figure to the right shows the same comparison for the angle algorithm and diaries in BBVS. TST: total sleep time

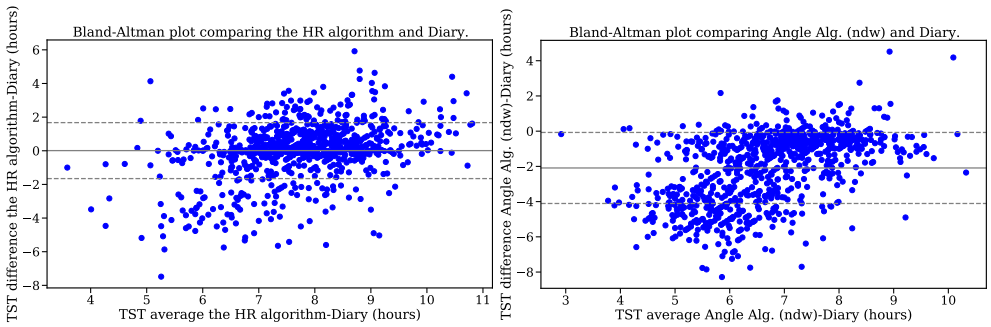
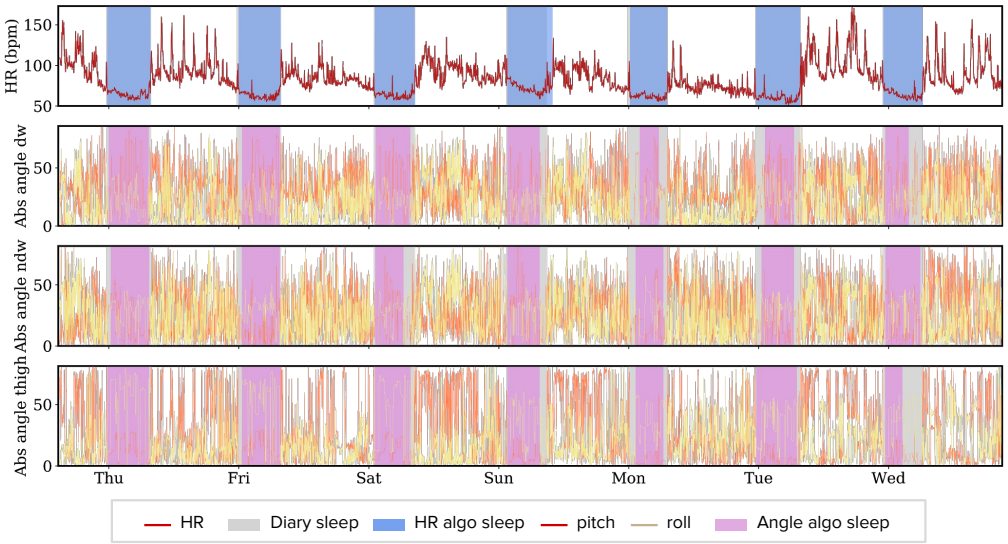


Figure 4.5 Example participant (chosen at random), showcasing estimated sleep through the heart rate sleep window algorithm, sleep diary sleep onset and offset and angle changes for both wrists and the thigh accelerometers. The algorithm picks up subtle sleep regularity differences at a participant level. This approach overlaps more closely to the sleep diary than any of the accelerometer-based approaches. Notice for the angle change approach the algorithm is more effective on the non-dominant wrist accelerometer than on the dominant wrist or thigh accelerometer for most nights. TST: total sleep time



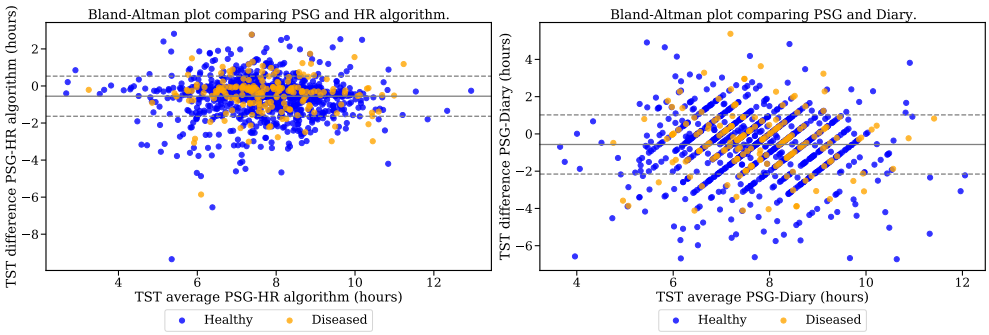
4.4.2 Evaluation of the algorithm in the MESA study

In MESA, we evaluated our algorithm against polysomnography for the full population, the subset of the population that were deemed healthy sleepers and those that were diagnosed with sleep disorders, a full breakdown of the results is presented in Table 4.3. Further, we evaluated the performance of sleep diaries against polysomnography in this study as well. Modified Bland-Altman plots for both the evaluation of our method against PSG and the diaries are presented in Figure 4.6.

Table 4.3 **Results for the MESA dataset.** Both HR algorithm and sleep diaries are evaluated against PSG. Results are also shown for the subset of healthy participants and participants with sleep disorders.

Sleep parameter	**Metric**	**HR algorithm** Value (mean ± 95% CI)	**Sleep Diary** Value (mean ± 95% CI)	**p-value**
Full cohort (n = 1154)				
Total sleep time	Time difference (min.)	-33.15 ± 0.12	-34.04 ± 0.18	0.756
	MSE	0.10 ± 0.00	0.13 ± 0.00	< 0.00
	Cohen's kappa	0.66 ± 0.00	0.62 ± 0.00	< 0.00
Sleep onset	Time difference (min.)	23.39± 0.10	6.25 ± 0.11	< 0.00
Sleep offset (Wake Up)	Time difference (min.)	-9.76 ± 0.08	-27.79 ± 0.16	< 0.00
Healthy participants (n = 965)				
Total Sleep Time	Time difference (min.)	-27.46 ± 0.28	-23.75 ± 0.42	0.605
	MSE	0.10 ± 0.00	0.13 ± 0.00	0.004
	Cohen's kappa	0.65 ± 0.00	0.60 ± 0.00	0.073
Sleep onset	Time difference (min.)	21.68 ± 0.23	6.92 ± 0.25	0.002
Sleep offset (Wake Up)	Time difference (min.)	-5.78 ± 0.20	-16.84 ± 0.37	0.084
Participants with sleep disorders (n = 189)				
Total Sleep Time	Time difference (min.)	-34.27 ± 0.13	-36.05 ± 0.19	0.565
	MSE	0.10 ± 0.00	0.13 ± 0.00	< 0.00
	Cohen's kappa	0.66 ± 0.00	0.62 ± 0.00	< 0.00
Sleep onset	Time difference (min.)	23.73 ± 0.11	6.12 ± 0.12	< 0.00
Sleep offset (Wake Up)	Time difference (min.)	-10.54 ± 0.09	-29.93 ± 0.17	< 0.00

Figure 4.6 Modified Bland-Altman plot for MESA. Modified Bland-Altman plot on the left shows the TST differences (delta) between the HR algorithm and PSG in the y-axis and the x-axis shows the TST average for every participant. The figure to the right shows the same comparison for the sleep diaries and PSG in MESA. Further, healthy participants are color coded in blue for both plots and participants that were diagnosed with sleep disorders are shown in orange.



The results in MESA reflect that our HR algorithm yields better performance to that of sleep diaries in terms of MSE for the full population (0.10 vs 0.13 MSE) as well as for the subset of the population with sleep disorders and healthy participants (0.10 vs 0.13 MSE). For all three analysis the best quantile was 0.85, likely due to the short amount of wake and active time in the recordings. The time differences in total sleep time are slightly less for the sleep diary (-34.04 versus -33.15 in the full population). Interestingly, while the time difference for sleep onset was around 20 minutes for the algorithm, it was only around 6 minutes for the sleep diary. In contrast, the algorithm approach fared much better at inferring sleep offset (between

-5 and -10 minutes of time deviation) whereas the diary underestimated sleep offset by almost 30 minutes. Finally, the algorithm approach yielded a stronger Cohen's kappa for all three sub-analysis than the sleep diary.

4.4.3 Evaluation of the algorithm in the PhysioNet Apple Watch Polysomnography study

Our algorithm was applied to data obtained from a commercial, readily available wrist-worn wearable and evaluated against gold-standard measures of sleep obtained with PSG. In this cohort, we evaluated both the HR algorithm and the angle change approach given the presence of triaxial accelerometry. The HR algorithm resulted in an MSE of 0.06 while the wrist-based angular change approach yielded an MSE of 0.12. Similar to MESA, the best performing quantile threshold was 0.8. This high quantile is likely due to the nature of the evaluation protocol consisting of concurrent PSG and wearable without much out of bed activity. Total sleep time deviation was of -23.23 minutes for the HR approach and 44.39 for the angle change approach. Sleep onset time deviation was of 15.12 minutes for the HR approach and of -21.77 for the angle change approach, while the difference was of -8.10 and 22.61 for sleep offset. However, Cohen's kappa was slightly lower for the HR approach (0.67) than for the angle change algorithm (0.71). These results are summarized in Table 4.4.

Table 4.4 **Results for the PhysioNet Apple Watch dataset.** The table presents results for both the HR and angle change algorithm for total sleep time, sleep onset and sleep offset in the PhysioNet Apple Watch dataset. ndw: Non-dominant Wrist

Sleep parameter	**Metric**	**HR Algorithm** Value (mean ± 95% CI)	**Angle change algorithm (ndw)** Value (mean ± 95% CI)	**p-value**
	Time difference (min.)	-23.23 ± 0.42	44.39 ± 1.30	0.003
Total sleep time	MSE	0.06 ± 0.00	0.12 ± 0.00	0.191
	Cohen's kappa	0.67 ± 0.00	0.71 ± 0.00	0.656
Sleep onset	Time difference (min.)	15.12 ± 0.16	-21.77 ± 0.96	0.021
Sleep offset (Wake Up)	Time difference (min.)	-8.10 ± 0.39	22.61 ± 1.00	0.058

4.4.4 Evaluation of the algorithm in the MMASH study

Our final set of evaluations took place in the MMASH cohort, which included both HR and triaxial accelerometer data recorded continuously for full-day periods. As in BBVS, we evaluated both the proposed HR approach and the angle change approach against detailed sleep diaries. We found that the results validated the findings from BBVS, resulting in the same optimal quantile (0.35) yielding an MSE of 0.07 and total sleep time difference of -14.08 minutes, with a cohen kappa of 0.85 for the HR approach. On the other hand, the angle change approach resulted in an MSE of 0.08 and cohen kappa of 0.83, but the total time deviation was substantially worse, yielding a total sleep time difference of -60.17 minutes. Full results for the MMASH cohort are presented in Table 4.5.

Table 4.5 **Results for the MMASH dataset.** The table presents results for both the HR and angle change algorithm for total sleep time, sleep onset and sleep offset in the MMASH dataset. ndw: Non-dominant Wrist

Sleep parameter	Metric	HR Algorithm Value (mean ± 95% CI)	Angle change algorithm (ndw) Value (mean ± 95% CI)	p-value
Total sleep time	Time difference (min.)	-14.08 ± 1.26	-60.17 ± 1.18	0.002
	MSE	0.07 ± 0.00	0.08 ± 0.00	0.344
	Cohen's kappa	0.85 ± 0.00	0.83 ± 0.00	0.366
Sleep onset	Time difference (min.)	-15.75 ± 0.68	15.13 ± 0.62	0.013
Sleep offset (Wake Up)	Time difference (min.)	-29.83 ± 0.97	-45.04 ± 1.08	0.022

4.5 Discussion

Objective and unobtrusive measurement of sleep in large, free-living populations at scale will help facilitate epidemiological investigations powered to explore the relationships between sleep, physical behaviours and disease. Concurrently, the rapid growth and adoption of commercial grade wearable devices offers a unique opportunity for the objective monitoring of sleep at scale. However, most commercial devices use algorithms that are not open-source or do not report thorough validation against gold-standard measures. Similarly, conventional algorithms tend to rely on device specific metrics, such as counts, requiring extensive adaptation for each device and cohort tested, as well as a predefined search window through expert annotations or sleep diaries. This often renders evaluation across devices and without sleep diaries futile.

Here we introduced a device agnostic algorithm that exploits the HR sensing capabilities present in most modern wearable devices. We presented an algorithm based on a personalized HR feature that allows detection of sleeping windows under free-living conditions. The proposed method relies on the well-established changes in HR that occur when individuals transition from wake to sleep [413]. Hence, it is able to infer sleep on individuals regardless of fitness level or illness and can be used amongst shift workers who exhibit sleep episodes outside of the night period. These qualities may be particularly relevant when evaluating sleep in populations with fragmented sleep, in countries where sleep timing changes due to seasonality or where cross-cultural sleep differences are observed [414]. The value of this approach lies in the fact that it is device-agnostic, does not require sleep diary or questionnaire data and adapts to inter- and intra-individual (day-to-day) variability, allowing for accurate and reliable sleep window labeling.

We evaluated our HR-based algorithm in four cohorts: BBVS, MMASH, PhysioNet and MESA. Both BBVS and MMASH include free-living HR, movement and sleep diary data for multiple days. By contrast, PhysioNet and MESA provide lab-based HR data and gold-standard PSG. Our aim was to evaluate the algorithm's performance in free-living conditions in the first two cohorts and compare it to existing measures that could be leveraged in these cohorts, whilst using the last two cohorts to evaluate its validity against gold-standard measures. Further,

through this process, we aimed to identify the range of parameters (Q, L, G) that produce the best results in free-living conditions, allowing for application and deployment in the absence of any ground truth.

For the first evaluation in the BBVS study, we found that the proposed method performed strongly in free-living conditions, with an average time deviation for total sleep time compared to non-habitual sleep diaries of -0.6 minutes. In this study, we performed optimal parameter search using both full day measures of HR as well as night-only measures to analyze how the availability of sensor data or the design of the experiment affect the choice of best parameters. The parameter search for the optimal MSE was performed based on quantile, window merge and window length values and are presented in Figure 4.7. We found that the optimal parameters for this cohort were 0.35 for the quantile (Q) and 30 minutes for the window length (L) and time merge block of 120 minutes (G). This resulting optimal quantile also makes intuitive sense as it represents about 8h, which is around the expected time spent sleeping for most individuals.

The algorithm performed better at detecting sleep offset (wake up), than sleep onset, yielding a time difference of 0.54 and 1.14 minutes respectively. This may be due to the fact that the sleep logs for validation of sleep onset and offset, while more detailed than traditional sleep diaries, rely on self-report and may not be wholly accurate. While sleep offset is relatively straightforward to annotate as most people wake up with alarm clocks, the exact time of sleep onset cannot be recorded, and is prone to measurement bias, if attempted at the time, or recall bias, if filled in the next day. Thus, the quality of self-reported sleep may vary based on the sleep onset latency of each participant for each given night. Nevertheless, the performance of the method across a diverse population and multiple nights of recording showcases its potential for free-living applications.

Finally, in the BBVS cohort, we evaluated the performance of an angle change-based algorithm inspired by previous work [390, 113] leveraging the multiple accelerometers available to evaluate postural changes. We found that this approach is valuable, but the results were more modest than those of our proposed method, yielding a total sleep time MSE of 0.10 and a time deviation of 125.37 minutes for the non-dominant wrist device. We also found that using the combined pitch and roll approach versus only the z-angle did not significantly alter the results. In sum, while valuable, the angle change approach performed significantly worse than our HR-based algorithm in the BBVS cohort. These results suggest that when HR is available, it should be used in preference, but triaxial accelerometry is a valuable second option in the absence of HR.

The algorithm was also evaluated in the MESA cohort, a large, diverse population where gold-standard PSG sleep measures through PSG were available, alongside self-reported sleep (through sleep diaries). In MESA we optimized the method to minimize MSE and additionally evaluated it in subsets of population with and without sleep disorders, yielding the results

reported in Table 4.3. In MESA, the deviation of total sleep time versus gold-standard measures of sleep was -33.15 minutes and MSE of 0.10 for the full population, whereas the same comparison between PSG and sleep diaries yielded a total sleep time deviation of -34.04 minutes and MSE of 0.13. This shows that our HR-based method can reliably and objectively monitor sleep in the absence of PSG and performs better than sleep diaries. It is also worth noting that the HR approach was significantly better at detecting sleep offset (wake up) with a time difference versus PSG of -9.76 minutes compared to the -27.79 of the diaries. These results further highlight that our algorithm can be used in the absence of sleep diaries and also shows superior performance in terms of MSE to conventional, habitual sleep diaries in this large cohort. Furthermore, the comparable results for the analysis carried out in the subset of the cohort with formally diagnosed sleep disorders point to the fact that our method may also be valuable when monitoring sleep in people suffering from these conditions. To the best of our knowledge, this is the first study that conducts these types of sensitivity analyses on a subset of disease subjects to show the validity of the proposed method in individuals who suffer from sleep disorders. Future work should carry more through validation on a larger population sample with sleep disorders to confirm these findings.

For the MESA cohort, recording in laboratory conditions and for less than 24 hours produces an HR ECDF that is different in shape from free-living conditions. The scarcity of non-sedentary activities means that the optimum quantile (Q) threshold needs to be higher to preserve its sensitivity at detecting sleep versus sedentary periods (optimum quantile of 0.85). This could have potentially constrained the performance of our method in this population. Future studies should explore continuous recording of HR during the day paired with full-PSG to test the validity of our method in a similar setting to that of BBVS or MMASH.

We further examined the performance of the HR algorithm in the PhysioNet Apple Watch cohort that recorded concurrent Apple Watch data and a night of PSG. This study followed a similar experimental protocol to that of the MESA study. In this cohort, the HR algorithm yielded an MSE of 0.06 and a time-deviation of -23.23 minutes when compared to gold-standard measures of sleep through PSG. Similar to MESA, the optimal quantile (Q) was quite high (0.8, with 0.85 producing the same results), which is likely due to the nature of the experimental setting where the Apple Watch recording period only started when the PSG was setup. Nevertheless, these results showcase the potential of the method in commercial-grade wearable devices that obtain HR through PPG. We also examined the angle change approach in this cohort, with this method performing less well than it did in BBVS, yielding an MSE of 0.12 and a total sleep time deviation of 44.39 minutes.

Finally, we validated our method in the MMASH cohort, where free-living HR, movement and sleep diary data was available. In this cohort, our algorithm's performance was optimal using a 0.35 quantile (Q) and window length (L) of which is the same result obtained for BBVS. Our HR approach yielded an MSE of 0.07 and a total sleep time deviation of -14.08. The angle change approach resulted in an MSE of 0.08 and total time deviation of -60.17. Our results in

MMASH confirm the strong validation performance we observed in BBVS when using our algorithm in free-living conditions across multiple days of recording. Based on the results obtained in BBVS which were then validated in MMASH, in free-living conditions a quantile (Q) of 0.35 lead to the best MSE results. Similarly, we recommend using a window length (L) in the range of 30-45 minutes and a merge block (G) in the range of 60-240 minutes as summarized in the pseudocode 1 formulation of our approach. These values can be used as priors for our algorithm an can be further fined tuned in the presence of ground truth.

One important limitation of the BBVS and MMASH studies is that they did not include PSG-derived ground truth sleep annotations. Although an ideal experimental protocol would have multiple days of PSG and free-living wearable sensor data, detailed sleep logs allowed us to evaluate the algorithm across more than one or two nights, showcasing the strength of our method in discerning both inter- and intra-individual variability. Similarly, the accelerometers included in these studies offer an important perspective on postural changes and how they compare to our proposed approach. Moreover, in ideal circumstances, HR for the full day would have been available in both the MESA and PhysioNet cohorts, optimizing the results of our approach by having exposure to non-sedentary wake behaviors. However, the results in these two datasets showcase the validity of our approach even under constrained laboratory conditions.

Future work should explore the robustness of the HR-based algorithm in cohorts such as inpatients. As the algorithm relies on HR signals already monitored continuously for other medical purposes, no additional accelerometer sensor would be required. Accurately labeling sleep in inpatients is challenging due to other factors that influence the HR ECDF, such as limited mobility, fever, medication, physiological and psychological stress, drug and alcohol use and cardiovascular conditions. However, objectively monitoring sleep without additional obtrusion could help improve sleep quality during hospital stays, which is a challenge for most patients [415], and hence promote both healing and patient satisfaction. Moreover, optimization of the angle change approach should be explored such that it can be used more reliably in the absence of HR sensor data, in this investigation we limited our evaluation of this method to the original parameters reported [390]. Parameter optimization could yield more generalizable and stronger outcomes for this approach. Finally, our method could be used in collaboration with some of the well-established activity-based approaches where multimodal settings are present. For instance, using conditional programming traditional methods could complement our approach in the detection of awakenings and assist in the derivation of conventional and novel sleep metrics.

Remark: Overall, our work highlights the potential of HR to detect the sleeping window not only in research and clinical contexts, but also in ecologically valid free-living conditions, enabling the objective monitoring of sleep in large-scale populations without PSG labels or sleep diary guidance. The low effort involved in collecting and analysing objectively inferred

sleep data coupled with low exclusion rate due to technical issues, missed diary entries or dropout would likely result in larger and more diverse study cohorts, as well as facilitating long-term objective data collection. For instance, few studies have been able to properly test the longitudinal, and likely synergistic, association between sleep quality and disease. Where this has taken place, sleep data is often collected through questionnaires [416] or with short, arbitrary follow-up periods. These studies could have missed long-term trends that significantly influence health status over months or years.

In sum, our proposed method was shown to accurately infer sleep in both free-living and laboratory conditions without the need for sleep diaries. As highlighted by Depner and colleagues [417], our analysis and evaluation will help enable the translation of findings from laboratory-based sleep studies into large-scale cohort studies and clinical trials, by providing an objective, device agnostic method to monitor sleep without the need for sleep diaries.

Chapter 4: Supplementary

Table 4.6 **Sleep Disorder Population details for the MESA study**. The MESA study allowed us to evaluate our method in a population which included sleep disorders with roughly the same prevalence as that of in the general population.

Total subjects	Healthy	Sleep apnea	Insomnia	Restless Legs Syndrome
1743	1469 (84.3%)	132 (7.6%)	109 (6.2%)	78 (4.5%)

Table 4.7 **Comparison of angle algorithm performance for the BBVS dataset by the limb on which the device was worn.** All participants wore devices on their dominant (dw) and non-dominant (ndw) wrist as well as on their thigh. The best performance metrics were obtained for the non-dominant wrist device, but thigh wearables gave the least time differences overall in terms of total sleep time (TST), sleep onset and offset.

Sleep parameter	**Metric**	**Angle change algorithm (ndw)** Value (mean ± 95% CI)	**Angle change algorithm (dw)** Value (mean ± 95% CI)	**p-value** ndw-dw	**Thigh** Value (mean ± 95% CI)	**p-value** ndw-thigh
Total sleep time	Time difference (min.)	125.37 ± 0.26	124.50 ± 0.26	0.814	118.68 ± 0.27	0.126
	MSE	0.10 ± 0.00	0.10 ± 0.00	0.032	0.10 ± 0.00	0.002
	Cohen's kappa	0.76 ± 0.00	0.76 ± 0.00	0.039	0.75 ± 0.00	0.007
Sleep onset	Time difference (min.)	-60.17 ± 0.21	-58.19 ± 0.20	0.562	-54.66 ± 0.24	0.175
Sleep offset (Wake Up)	Time difference (min.)	65.20 ± 0.20	66.32 ± 0.20	0.710	64.01 ± 0.23	0.756

Algorithm 1 Method to estimate sleep periods based on Heart Rate.

Input: W - Wearable Data with Heart Rate Data
Q - Quantile Value (Default: 0.40)
L - Minimal window length (Default: 45)
G - Maximal gap interval in minutes to merge sleep windows (Default: 60)

Output: Sleep window inferences

Function `HR_Algorithm`**:**

```
/* Split data into "experiment days" from 3pm-to-3pm.                    */
W_D = split_days(W, time=15)
for d ∈ D do
    /* Extract HR data from wearable device                              */
    HR = get_HR(W_d)
    /* Calculate quantiles for day                                       */
    HR^Q = calculate_quantile(HR, Q)
    /* Get sequences that HR < HR^Q                                      */
    SleepArrays = get_sleep_sequences(HR, HR^Q)
    /* Keep only sequences larger than W.                                */
    for sleepArray ∈ len(SleepArrays) do
        if lengthInMinutes(sleepArray) < L then
            remove(sleepArray)
        end
    end
    /* Merge Sequences if gap between them is smaller than G             */
    for i ∈ len(SleepArrays) do
        if get_gap(SleepArray_i, SleepArray_{i+1}) < G then
            merge(SleepArray_i, SleepArray_{i+1})
        end
    end
    /* Select Limits of merged Sleep Window                              */
    for limit ∈ (onset; offset) do
        Select searchWindow = (limit − 240epochs; limit + 60epochs)
        HR Vol = get_rolling_std_dev(searchWindow, window = 10 epochs)
        From searchWindow select epochs where HRVol ≥ 6 beats per min and add to
          highVolatilityList
    end
    /* Define final Sleep Window                                         */
    if onset select last epoch from highVolatilityList then
        Overwrite limit as last epoch
    end
    if offset select first epoch from highVolatilityList then
        Overwrite limit as first epoch
    end
end
return
```

Figure 4.7 Mean Square Error (MSE) results for Biobank Validation Study (BBVS) using the full-day Empirical Distribution Function method to detect sleep windows. The MSE was calculated through evaluation against sleep diary. The Y axis represents the quantiles tested for the analysis while the X axis are the window lengths. The optimal combination found through this search was a quantile of 0.35, time merge block of 120 minutes and a window length of 30 minutes, yielding an MSE of 0.06 in the BBVS study.

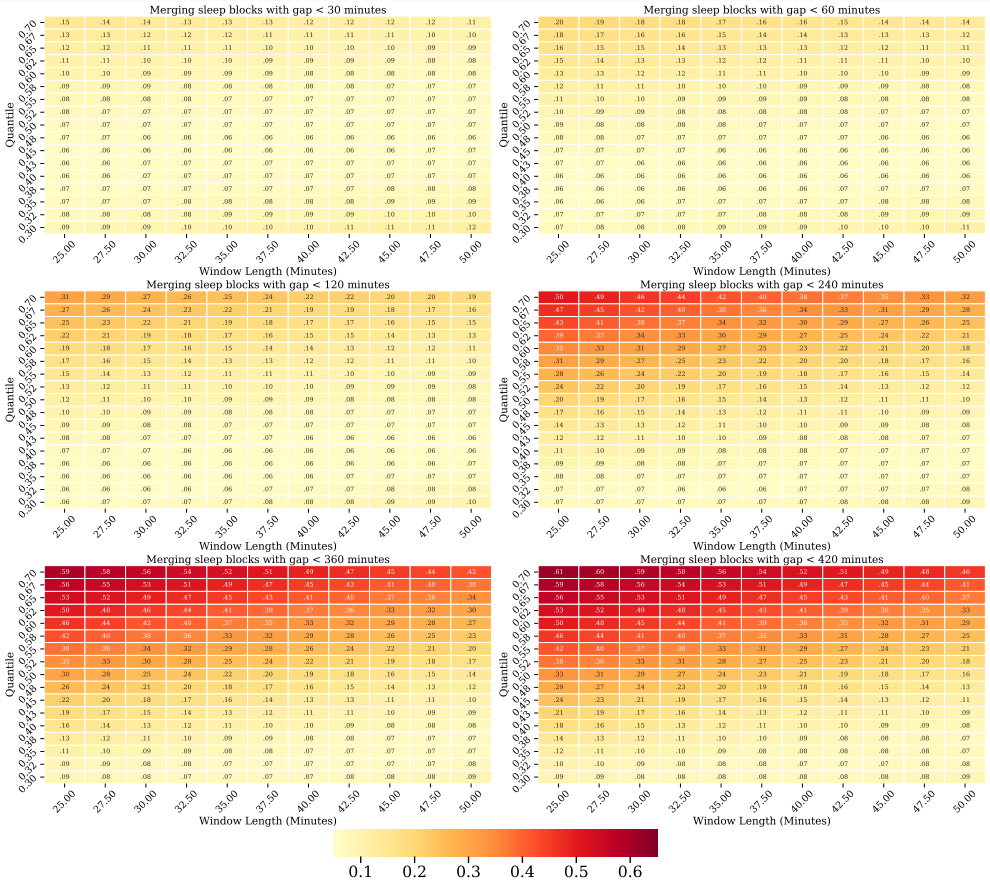


Table 4.8 **Results of applying the HR algorithm on the BBVS dataset for both full-day and night-only data.**

Sleep parameter	**Metric**	**HR algorithm (Full day)** Value (mean, 95% CI)	**HR algorithm (Night only)** Value (mean, 95% CI)	**p-value**
Total sleep time	Time difference (minutes)	-0.60 ± 0.21	-8.47 ± 0.20	0.050
	MSE	0.06 ± 0.00	0.06 ± 0.00	< 0.00
	Cohen's kappa	0.86 ± 0.00	0.87 ± 0.00	< 0.00
Sleep onset	Time difference (minutes)	1.14 ± 0.20	0.48 ± 0.16	0.832
Sleep offset (Wake Up)	Time difference (minutes)	0.54 ± 0.16	-7.98 ± 0.15	0.004

Figure 4.8 Mean Square Error (MSE) results for Biobank Validation Study (BBVS) using the night-only Empirical Distribution Function method to detect sleep windows. The MSE was calculated through evaluation against sleep diary. The Y axis represents the quantiles tested for the analysis while the X axis are the window lengths. The optimal combination found through this search was a quantile of 0.55, time merge block of 360 minutes and a window length of 42.5 minutes, yielding an MSE of 0.06 in the BBVS study.

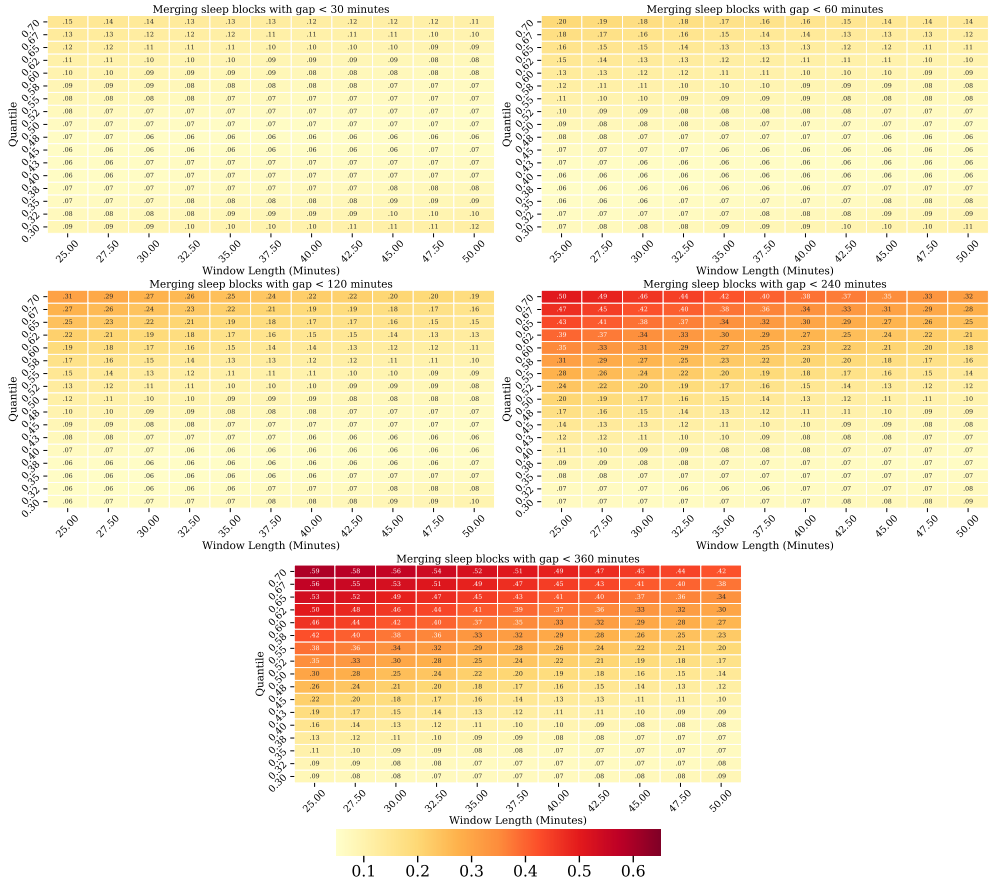
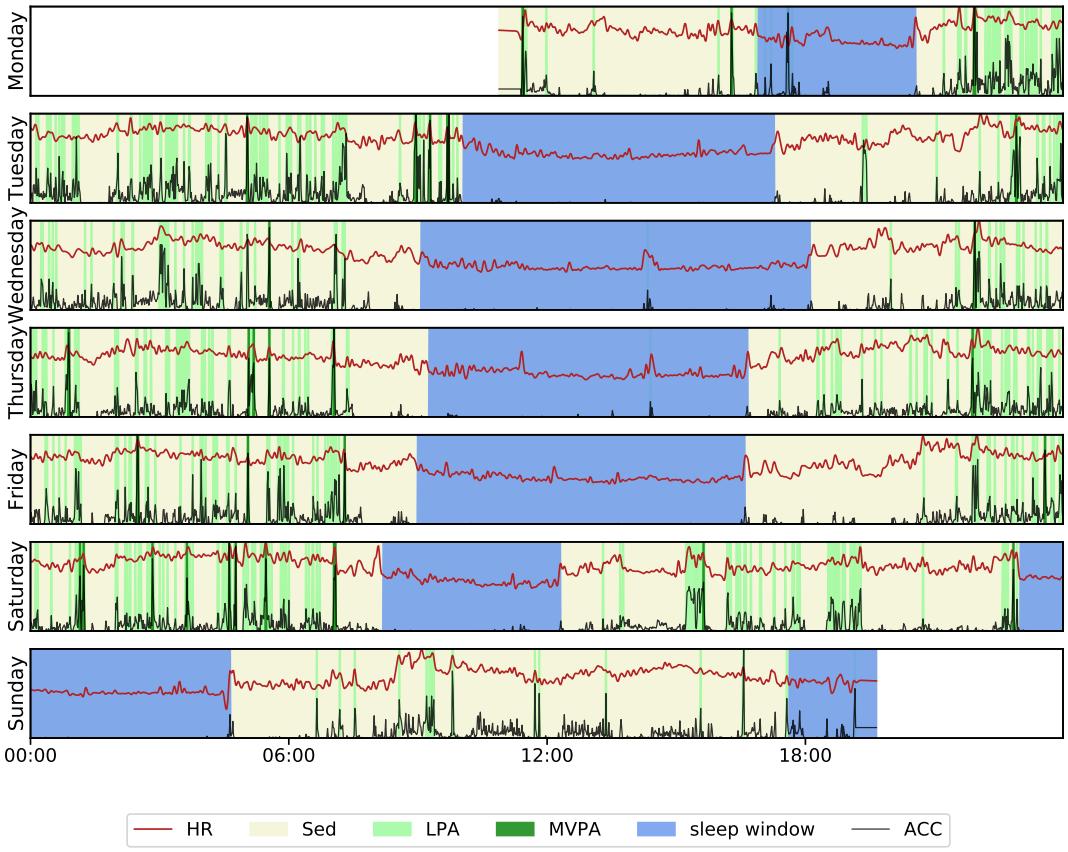


Figure 4.9 Applying the HR sleep algorithm on a shift worker. The free-living trace shows the subtle changes for day of the week picked up by the algorithm, with 2 sleep windows detected on Saturday, when they were not at work during the night. HR: Heart Rate; Sed: Sedentary; LPA: Light Physical Activity; ACC: Acceleration.



CHAPTER 5

UNMASKING PHYSICAL BEHAVIOURS IN LARGE POPULATION STUDIES THROUGH POSTURAL ANALYSIS DERIVED FROM WRIST-WORN ACCELEROMETRY

Publications

This study has been published:

Perez-Pozuelo, I., White, T., Westgate, K., Wijndaele, K., Wareham, N. J., Brage, S. (2019). Diurnal profiles of physical activity and postures derived from wrist-worn accelerometry in UK adults. Journal for the Measurement of Physical Behaviour, 1(aop), 1-11..

Contributions

I planned this project and devised the analysis plan in collaboration with my supervisors. I processed the data, created the models, wrote the Python code, designed the analysis and wrote the manuscript and this chapter.

5.1 Summary

Background

Wrist-worn accelerometry is the commonest objective method for measuring physical activity in large-scale epidemiological studies. Research-grade devices capture raw triaxial acceleration which, in addition to quantifying movement, facilitates assessment of orientation relative to gravity. No population-based study has yet described the interrelationship and variation of these features by time and personal characteristics.

Methods

2043 UK adults (35-65years) wore an accelerometer on the non-dominant wrist and a chest-mounted combined heart-rate-and-movement sensor for 7 days free-living. From raw (60Hz) wrist acceleration, we derived movement (non-gravity acceleration) and pitch and roll (forearm) angles relative to gravity. We inferred physical activity energy expenditure (PAEE) from combined sensing and sedentary time from approximate horizontal arm-angle coupled with low movement.

Results

Movement differences by time-of-day and day-of-week were associated with forearm-angles; more movement in downward forearm-positions. Mean(SD) movement was similar between sexes $\sim$31(42)mg, despite higher PAEE in men. Women spent longer with the forearm pitched $>0^\circ$, above horizontal, (53% vs 36%) and less time at $<0^\circ$ (37% vs 53%). Diurnal pitch was 2.5-5° above and 0-7.5° below horizontal during night and daytime, respectively; corresponding roll angles were $\sim 0^\circ$ (hand flat) and $\sim 20^\circ$ (thumb-up). Differences were more pronounced in younger participants. All diurnal profiles indicated later wake-times on weekends. Daytime pitch was closer to horizontal on weekdays; roll was similar. Sedentary time was higher (17 vs 15 hours/day) in obese vs normal-weight individuals.

Conclusions

More movement occurred in forearm positions below horizontal, commensurate with activities including walking. Findings suggest time-specific population differences in behaviours by age, sex, and BMI. The findings of this work also inspired the approaches used to evaluate postural changes in the previous Chapter of this Thesis.

5.2 Background

Wrist-worn accelerometry has become a feasible option for the objective measurement of physical activity in large-scale epidemiological studies, such as Pelotas birth cohorts, the UK Biobank and Whitehall II [418, 3, 419]. Additionally, public adoption of consumer-grade wearable devices that include accelerometry has been increasing steadily in recent years [420, 421], with potential utility for public health research [39].

Accelerometers record a continuous time-series of data and recent advances in the technology and battery life allow for ubiquitous capture of raw accelerometer signals which have the potential to provide insights to interventional and epidemiological studies. Several features can be easily extracted from the acceleration signal, including the magnitude of movement and the orientation of the accelerometer with respect to gravity.

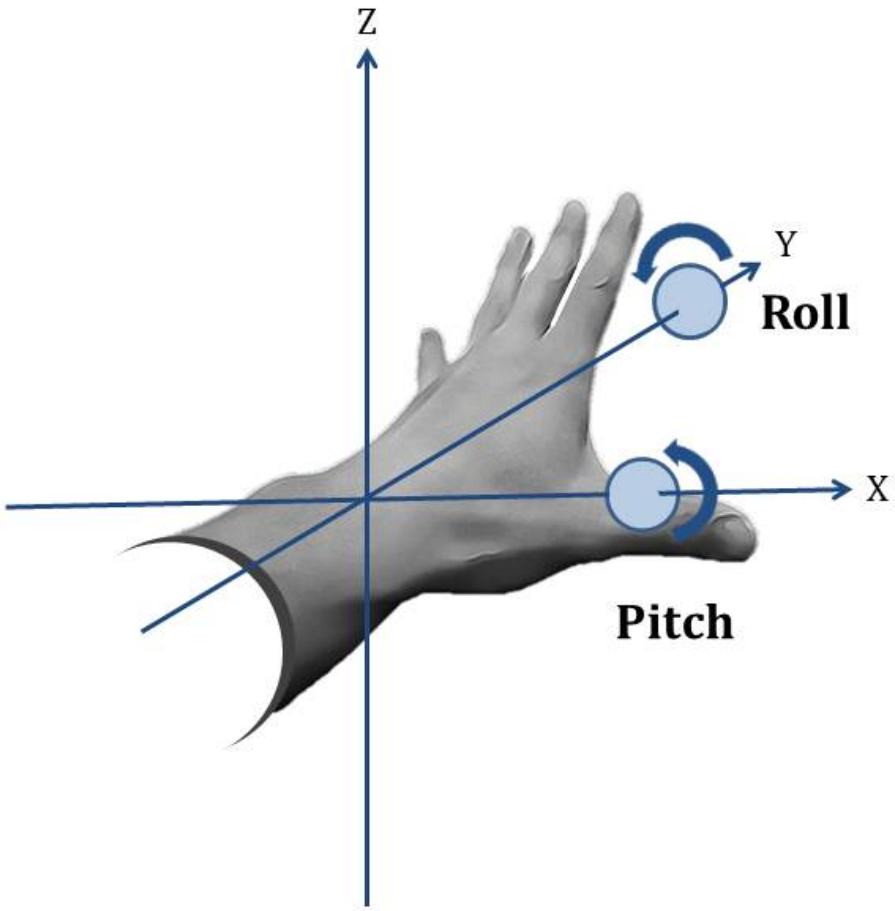
Chapter Significance: This chapter introduces the derivation of postural wrist measures using accelerometer data. The implications of this work relate to the possibilities associated to the use of these biomechanical inferences alongside traditional intensity based metrics like vector magnitude.

Previous research using wrist accelerometry has described variation in population physical activity expressed predominantly as the activity-related acceleration magnitude as explored in Chapter 1. For example, da Silva et al noted age and sex differences in three Brazilian birth cohorts from Pelotas assessed at the ages 6, 18, and 30 years of age [418], and Doherty et al described the unique diurnal patterns of physical activity by age group, documenting that the lower activity levels generally observed in older adults are particularly pronounced in the later hours in the day [3]. Magnitude-based measures of activity have also been related to health outcomes, such as body composition and fitness [422, 39].

Less attention has been given to the description of orientation-related measures of human behaviour using wrist-worn acceleromery, although they have been used in thigh-worn activity monitoring [423]. Pitch and roll angles are examples of well-defined, biomechanically relevant and easy-to-interpret signal features that describe device orientation. In Figure 5.1, we illustrate pitch and roll for an individual with an accelerometer placed on the left wrist, and axes aligned as shown. Body posture is by definition a description of angles of all segments of the body and is determined by their relationship with gravity, which can in theory all be measured but in practice usually are not in studies of free-living behaviour. However, as body segments are connected, and therefore range of motion is restricted, measurements and the variables that can be mathematically derived from them tend to be highly correlated [397]. This allows inferences from the measurement of one body site to be made on whole-body posture. For example, previous work has shown strong correlations between time spent sedentary inferred

from wrist accelerometry (by combining information on acceleration magnitude and pitch angle) and thigh accelerometry (r∼0.93) [424].

Figure 5.1 Schematic of forearm Pitch and Roll on participant with accelerometer on the left wrist, including axes alignment. Roll is defined by rotation around the y-axis, while Pitch is defined by rotation around the x-axis. (Note that axis labeling depends on study protocol and device specifications)



Sedentary behaviour can be defined as any waking behaviour that is characterized by an energy expenditure $\leq$ 1.5 METs while the subject is engaging in either sitting, lying or reclining postures [425]. People spend the majority of their time in sedentary behaviours, and the proportion of time spent sedentary increases as people age [426]. High volumes of sedentary behaviour have been associated with increased mortality and risk of developing chronic conditions [43, 427, 426, 428, 425]. This only seems to be eliminated by very high levels of moderate intensity physical activity (60-75 min per day, i.e. equivalent to double the amount currently recommended for adults [427]). However, most of this evidence base is based on self-reported sedentary and activity estimates which come with important methodological limitations and bias [429].

Consequently, objectively assessing sedentary behaviours, as well as characterizing different activities performed during daily living may be critical to inform public health recommendations. Traditionally, sedentary and active behaviours were characterized using such intensity derived measures from the accelerometer signal. Figure 5.8 provides a visual representation of triaxial wrist acceleration (top panels) during four common activities of lying, walking, sitting, and cycling, alongside derived pitch and roll angles (bottom panels), demonstrating clear differences between activity types. When assessing activity patterns, diurnal profiles of pitch and roll combined with movement intensity metrics may allow us to further understand how different postures relate to different activities and activity intensities.

In this chapter, we describe the distribution of forearm postures, acceleration, derived sedentary time and PAEE in a large cohort of UK adults (n=2043 participants). These analyses allow us to further understand the distribution of sedentary and active behaviours in the population and how this distribution may differ based on time of the day, sex, age, body mass index (BMI) and overall activity levels. Ultimately, the methodology developed for the work presented aims to help inform how changes in sedentary and active behaviours may impact energy expenditure.

5.3 Methods

5.3.1 Study population

The Fenland Study is an ongoing prospective cohort study of 12,435 men and women aged 35-65 years, designed to identify the behavioural, environmental and genetic causes of obesity and type-2 diabetes. As previously described in detail, participants attended one of three clinical research facilities in the region surrounding Cambridge, UK, and completed a series of physical assessments and questionnaires [396]. Exclusion criteria for participation in the study were: clinically diagnosed diabetes mellitus, inability to walk unaided, terminal illness, clinically diagnosed psychotic disorder, pregnancy or lactation. Following the baseline clinic visit, all participants were asked to wear a combined heart rate and movement sensor (Actiheart, CamNtech, Cambridgeshire, UK) for 6 consecutive days and nights, and a subsample of 2100 participants were asked to simultaneously wear a wrist accelerometer (GeneActiv, ActivInsights, Cambridgeshire, UK) on the non-dominant wrist. This subsample constitutes the sampling frame for the current analyses. Participants were excluded from this analysis if they had insufficient individual calibration (treadmill test-based) data, or had less than 72h of concurrent wear data (equivalent of 3 full days of recording). Given only very few participants were very severely underweight (BMI $\leq$ 15) in this subset of the Fenland study, they were also excluded, resulting in a total of 2043 subjects. All participants provided written informed consent and the study was approved by the University of Cambridge research ethics committee and performed in accordance with the Declaration of Helsinki.

5.3.2 Data Collection

5.3.2.1 Physical Activity Measures

The combined heart rate and movement sensor attached to the participant's chest, measured heart rate and uniaxial acceleration of the trunk in 15-second intervals [401]. The wrist accelerometer worn on the non-dominant wrist recorded triaxial acceleration at 60 Hertz. Participants were instructed to wear both waterproof monitors continuously for 6 full days and nights during free-living conditions, including during showering and while they were sleeping. During the baseline clinic visit, participants performed a ramped treadmill test to establish their individual heart rate response to a submaximal exercise test . These measurements produced calibration parameters that were used in a branched equation model of PAEE [430]. Heart rate data collected during free-living was pre-processed to eliminate potential noise [402], following which the branched equation model was applied to calculate instantaneous PAEE (J$\cdot\, min^1 \cdot kg^1$). This inference has been validated against intensity from indirect calorimetry [431, 432] and volume from doubly-labelled water in several populations [433], including

a sample of UK men and women in whom the technique was shown to explain 41% of the variance in free-living PAEE as well as no mean bias [434]. The wrist accelerometer data was processed using pampro, an open-source software package. The triaxial acceleration was auto-calibrated to local gravitational acceleration using a method described elsewhere [403]. Non-wear time was defined as time periods where the standard deviation of the acceleration in each of the three axes fell below 13mg for over an hour, inferring that the device was completely stationary [435]. When a non-wear period was detected, it was removed from the analyses. The magnitude of acceleration was calculated using Vector Magnitude (VM) (expressed in milli-g/mg) per sample:

$$VM(X,Y,Z) = \sqrt{(X^2 + Y^2 + Z^2)} \tag{5.1}$$

VM, or Euclidean Norm, can be interpreted as the magnitude of acceleration the device was subjected to at each measurement, which includes gravitational acceleration. Any potential noise component in the high-frequency domain was filtered out by a 20 Hertz low-pass filter. To isolate the movement-related acceleration, we also applied a high-pass Butterworth filter to the VM signal at 0.2 Hertz (therefore treating gravity as a low-frequency component) naming the resulting metric Vector Magnitude High-Pass Filtered (VM HPF, expressed in mg) [435, 39]. VM HPF is commonly used as a proxy of acceleration resulting from human movement, has high validity [37], and was the primary description of wrist movement in the following analyses. To isolate the gravitational acceleration for each axis, we applied a low-pass filter (0.2 Hertz) to each of the three axes (X, Y and Z). The residual acceleration signal can be interpreted as a measurement of the rotated gravitational field vector which can then be used to determine the accelerometer's pitch and roll orientation angles. Pitch and roll of the device were derived according to these formulae:

$$Pitch = \frac{tan^{-1}\left(\frac{Y}{\sqrt{X^2+Z^2}}\right) * 180}{\pi} \tag{5.2}$$

$$Roll = \frac{tan^{-1}\left(\frac{X}{\sqrt{Y^2+Z^2}}\right) * 180}{\pi} \tag{5.3}$$

As the monitor was mounted in such a way that the X-axis was aligned in anatomically opposite directions for left- and right-handed participants, we multiplied it by -1 for all left-handed participants who wore the monitor on their right wrists to align with the anatomical coordinate system defined above (examples of untransformed data shown in Figure 5.9). Consequently, positive pitch indicates upwards position of the arm (hand above elbow), while positive roll indicates the lateral (radial, thumb) side of the arm being higher than the medial (ulnar, pinky) side of the arm. Figure 5.1 illustrates these concepts. All derived signals were summarized to a common time resolution of one observation per hour. This window length was chosen since we were mostly interested in observing changes at a diurnal level, rather than variations within the hour. Using the combined-sensing measurements, participants were stratified by average

activity energy expenditure into three equal tertiles: lower active ($\leq$ 39 J $\cdot$ min^{-1} $\cdot$ kg^{-1}), medium (40-56 J $\cdot$ min^{-1} $\cdot$ kg^{-1}) and upper ($\geq$ 57 J $\cdot$ min^{-1} $\cdot$ kg^{-1}). These activity estimates were calculated for each participant for each day of the week and then averaged, allowing us to generate a activity-level stratification based on the individual weekly average. picture of changes in behaviour over the course of the week. Similarly, we calculated estimates of time spent in sedentary (i.e. sitting or reclining) by detecting bouts where wrist pitch (i.e. fore arm elevation) is $\geq$ 15 $^\circ$ below the horizontal, while wrist acceleration is minimal (VM HPF $\leq$ 47.61 mg)). This is based on principles from previously developed methodology which derives sedentary time estimates from wrist accelerometry data (i.e. sedentary sphere methodology [424]), as well as estimations of physical activity energy expenditure in free-living using wrist accelerometry [397]. The latter defined the acceleration threshold (VM HPF =47.61 mg) equivalent to 1.5 gross METs (PAEE=35.5 J$\cdot$ min^{-1} $\cdot$ kg^{-1}) as the cut-off for sedentary behaviour [39]. Data in lower latitudes, that is, less than -15° from the horizontal, suggest hanging of the arm, associated to standing behaviours and are hence not classified as sedentary time. Equally, if the mean levels (VM HPF) over a minute fell into the light, moderate or vigorous category, they were not classified as sedentary behaviour. Using the diurnal profiles derived from the cohort, we studied differences based on sex, age, activity levels, BMI and time of the day.

5.3.3 Statistical analyses

We computed descriptive statistics (mean, median, standard deviation, minimum, maximum and variance) for the participants in this analysis. We examined wear-time distributions using the Friedman test for time-of-day (00:00-05:59, 06:00-11:59, and so on in six-hour periods) and tested the differences in weekdays versus weekend days using Wilcoxon signed ranks. These tests were performed in men and women separately. Mean acceleration differences (VM HPF) were examined using ANOVA for time of the day and day of the week. Differences between men and women are shown by using box plots, providing information about the median, inter-quartile range, minimum and maximum. We analysed the differences between different BMI groups (underweight $\leq$ 18.5 kg/m^2, normal weight 18.5-24.9 kg/m^2, overweight 25-29.9 kg/m^2, obese 30-34.9 kg/m^2 and severely obese $\geq$ 35 kg/m^2) in both sexes based on pitch, roll, VM HPF and PAEE. Similarly, we conducted the analysis based on age group and PAEE levels. These summary statistics were computed at an hourly level after collapsing information derived on a fifteen-second time window. Furthermore, we tested for differences in time spent in sedentary time across the different BMI populations using 3-way ANOVA and adjusting for age and sex. Statistical tests were performed using Python (3.6.2) and Stata (v14, StataCorp, TX, USA).

5.3.4 Results

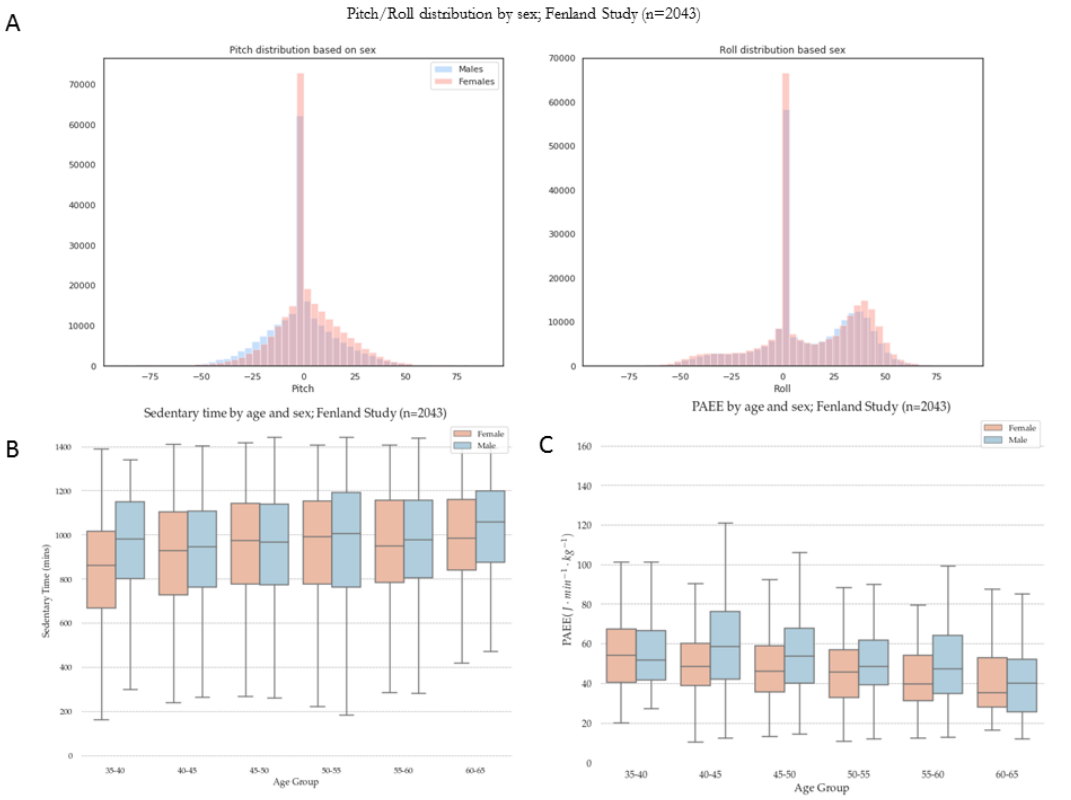
Among the 2043 participants, a total of 286,020 person-hours were included in our analysis, or an average of 5.8 days per participant. As shown in Table 1, PAEE was higher in men although both groups had large standard deviations. However, wrist movement was similar between genders but mean BMI was larger in men than in women for this cohort.

Table 5.1 **Characteristics of Participants by Sex (n=2043)** Values are means (standard deviations)

	Men	Women
N	953	1090
Age (years)	50.9 (7.3)	50.5 (7.1)
Height (m)	1.78 (0.07)	1.64 (0.06)
Weight (kg)	86.2 (14.1)	70.8 (14.3)
BMI (kg $\cdot$ m^2)	27.2 (4.2)	26.4 (5.3)
PAEE (J $\cdot$ $min^{-1} \cdot kg^{-1}$)	53.1 (21.9)	47.7 (19.1)
Wrist movement, VM HPF (mg)	31.7 (44.9)	31.1 (40.8)

Figure 5.2A shows pitch and roll distributions for men and women. There is higher occurrence of pitch and roll positions around 0° and the roll distribution is distinctly bimodal with an additional peak around 35°. Less common are extreme anatomical forearm positions e.g., arms up in the air, reflected by a pitch >60°, or the radial (thumb) side of the arm turned inwards and downwards as indicated by less roll data below -45°. Figure 5.2B and 5.2C shows the differences among different age groups for average sedentary time and PAEE respectively. PAEE declines with age in both men and women, and there is a tendency for the wrist measure of sedentary time to increase with age, showing a close inverse relationship between these two measures.

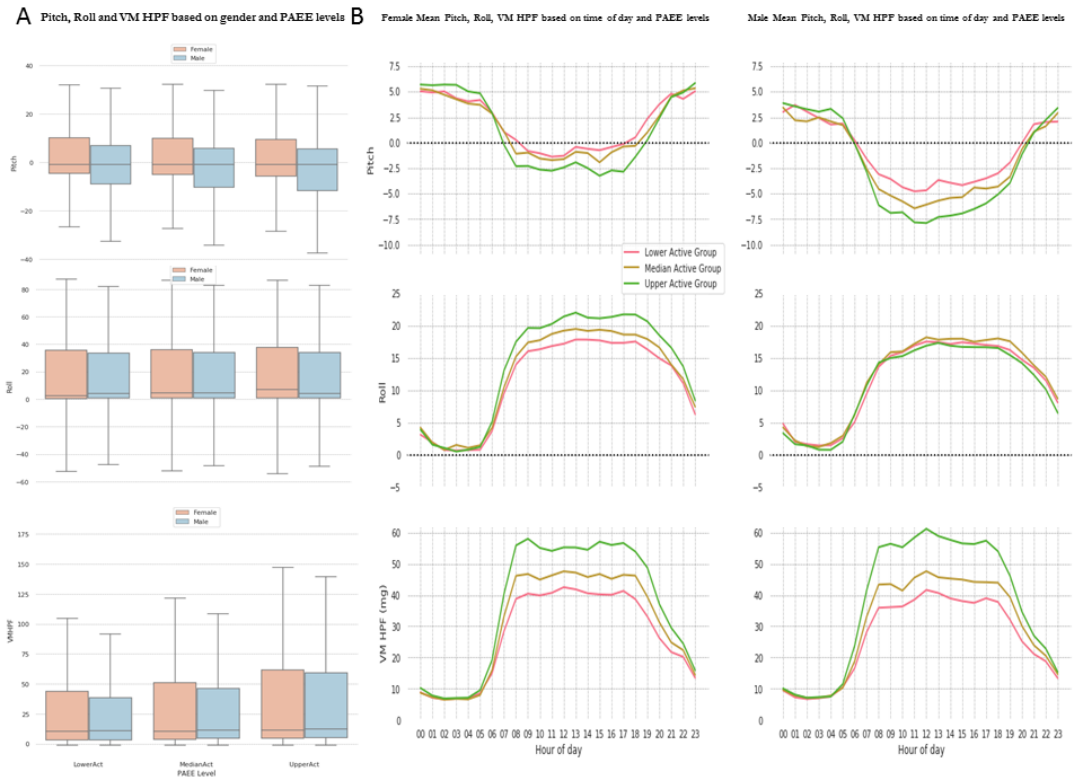
Figure 5.2 Pitch and roll (A) distribution among participants, and box plots for time spent sedentary (B) and PAEE (C) by age group and sex (n=2,043).



5.3.4.1 Relation between wrist movement and forearm postures, and physical activity energy expenditure

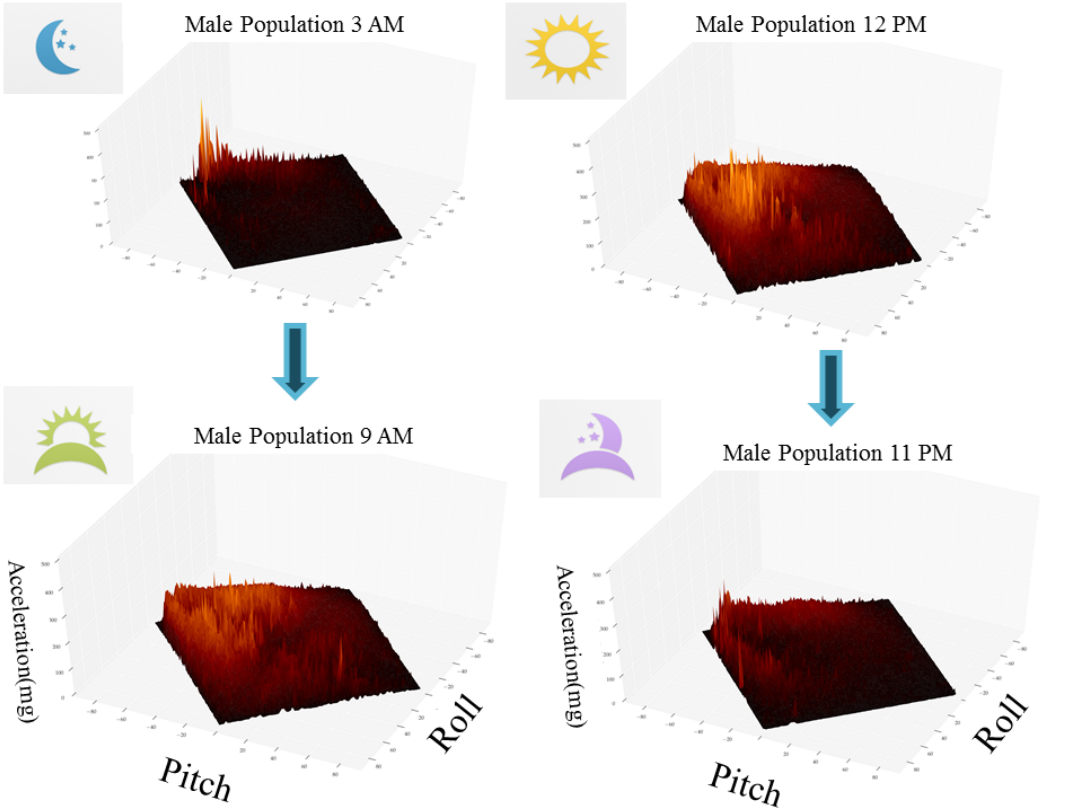
Figure 5.3 shows differences in wrist measures by tertile of physical activity energy expenditure; more active individuals spend more time in low-pitch (below horizontal) postures; less active participants tend to be spending more time in postures that suggest sedentary behaviours, such as sitting or reclining. Whilst roll angles differ by activity level in women, there is almost no difference between groups in men; differences in wrist movement, however, are very clear in both genders.

Figure 5.3 Pitch (top panels), Roll (middle panels), and Vector Magnitude High-Pass Filtered (VM HPF) by physical activity energy expenditure level (lower, medium, or upper) and gender (A), and diurnal profiles (hourly averages) by time of day in women and men (B).



Some of the most visually striking results regarding the role of posture on physical activity behaviours can be seen in the 3-dimensional time-lapse plots that appear on the online supplementary online material of this paper (see video) published on the Journal of the Measurement of Physical Behaviour (2019) [113]. A schematic representation of these time-lapses is presented in Figure 5.4 at four times of the day.

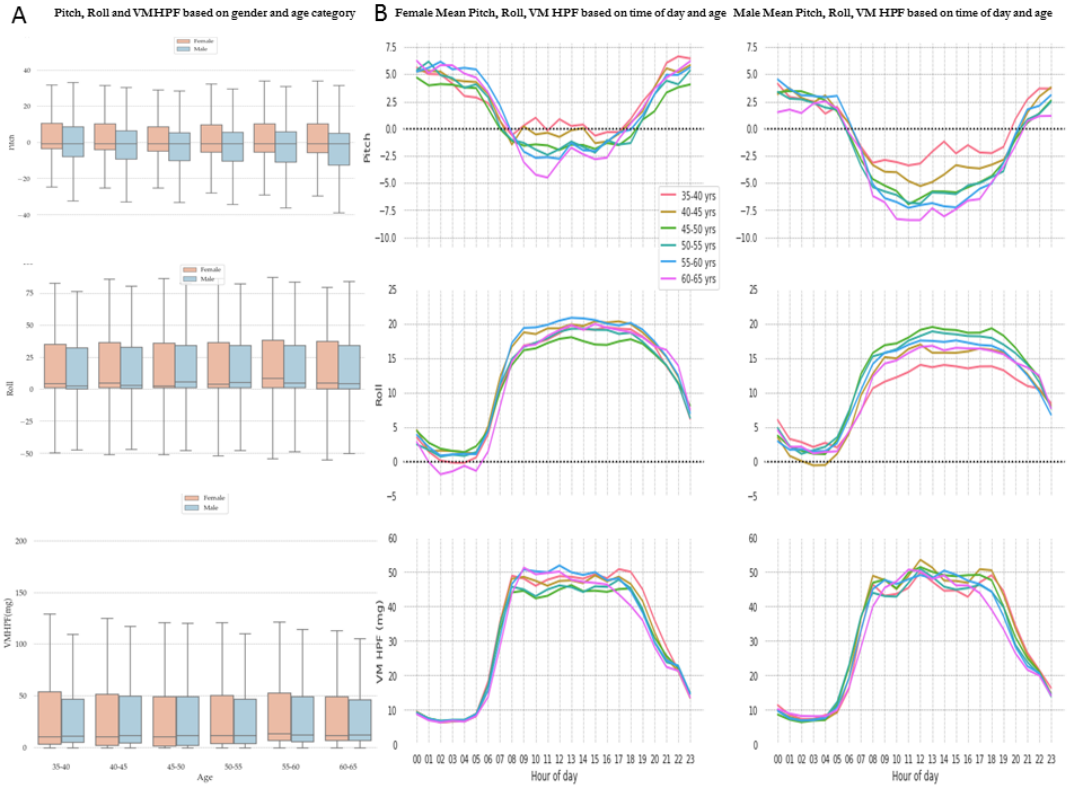
Figure 5.4 Schematic representation of time-lapse diurnal change in Pitch and Roll angular profiles and their associated acceleration signal (Vector Magnitude High-Pass Filtered, in mg). All plots have been normalized. (Figure derived from the male population of this analysis n=953).



5.3.4.2 Diurnal Profile Differences by sex and age

Figure 5.5 shows the distribution of pitch, roll, and movement intensity across the day, stratified by sex and age group. We observe differences between age groups within sex, but also differences between men and women within age groups. Most differences between men and women occur during the working hours (8 AM to 6 PM) of the day, with little differences at night although women generally keep their arms at slightly higher pitch throughout the 24 hours. Some of the biggest differences between age groups in both sexes happen during the early hours of the morning and late hours of the evening. Forearm angles differ more between age groups in men (lower pitch in older during working hours), and gender differences in pitch and roll profiles are most apparent among the 35-40 age group.

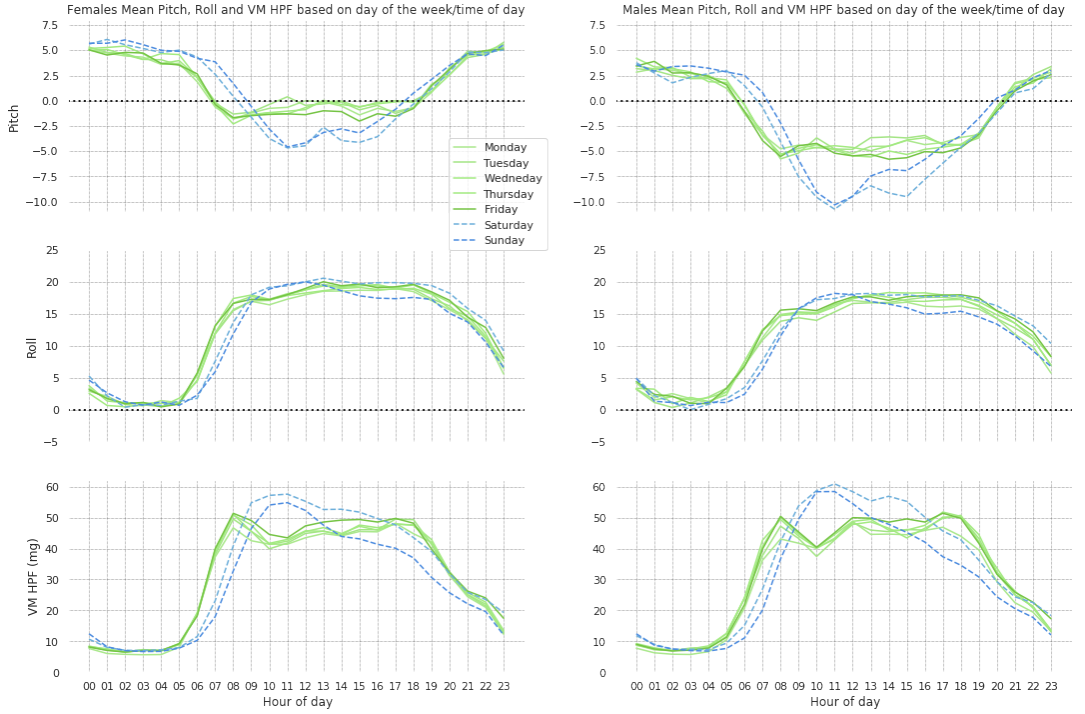
Figure 5.5 Pitch (top panels), Roll (mid panels), and Vector Magnitude High-Pass Filtered (bottom panels) profiles (hourly averages) by time of the day and age group (from 35–40 to 60–65 years old) in women (middle column) and men (right column). Left column (A) shows participant-level summary data.



5.3.4.3 Pitch and Roll Profiles Differ on Weekends versus Weekdays

Figure 5.6 shows average pitch, roll and movement intensity across the day, separately for each day of the week, and stratified by sex. The variation between weekdays at a population level is minimal, but they differ from the diurnal profiles at the weekend and particularly among sexes. A visible shift on weekend days towards later hours of the morning suggests a "later start" to the day, and later bed times on Friday and Saturday nights. The most extreme postural contrast are seen for pitch angles in men which reach the lowest level at the weekend (around -10°) in parallel to highest level of movement; pitch in women is also lower in the weekend but only to the weekday level of the men (around -5°) but with a similar level of movement as men.

Figure 5.6 Differences in Pitch, Roll, and Vector Magnitude High-Pass Filtered (hourly averages) based on day of the week (solid lines indicate weekdays , dashed lines indicate weekends) and time of the day in women and men.

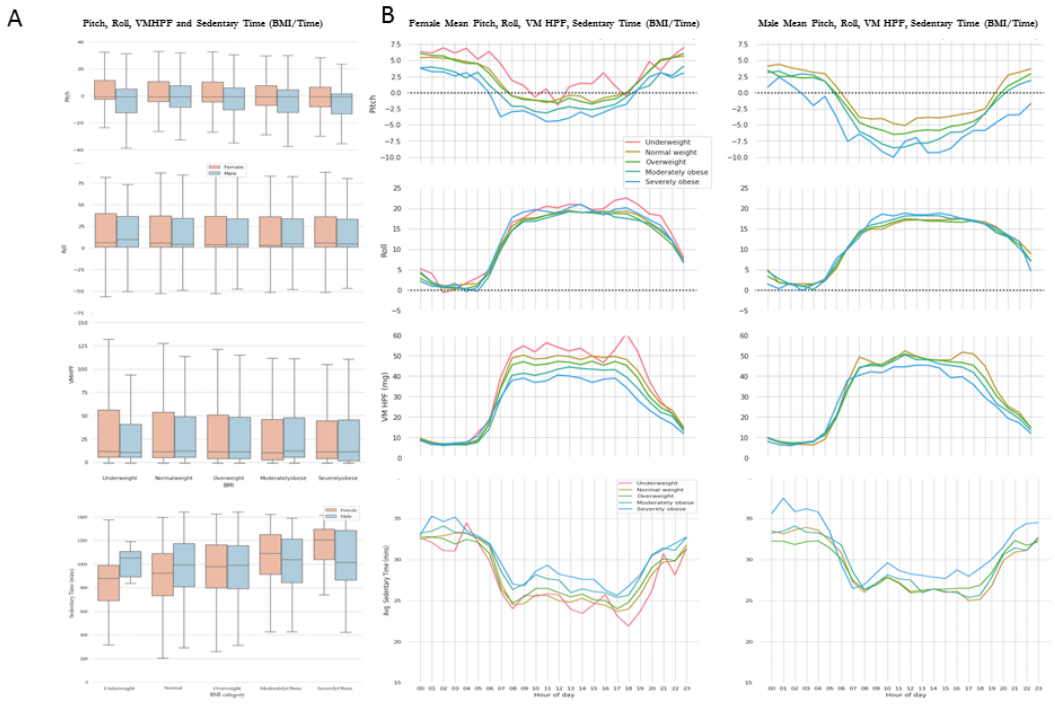


5.3.4.4 Wrist Accelerometry Profiles by Gender and BMI

The differences in mean VM HPF between the different BMI groups are striking with obese individuals moving considerably less than normal-weight but equally notable are differences in pitch and roll profiles (Figure 5.7). Differences among groups were more apparent in men than in women when considering the diurnal profile. Somewhat surprisingly, given higher movement is generally occurring at the lower pitch angles (Figure 5.3), overweight and obese individuals spend more time with their arms in this space but they just do not seem to move as much. The underweight women's pitch and roll profile are very different to that observed in the severely obese men, suggesting that the higher level of mean physical activity in this group is also related to a very different set of activities. These observations are supported by stark differences on the average time spent in sedentary behaviours stratified by sex and BMI category, where non-obese participants spent considerably less time in sedentary behaviours than obese participants, particularly women. Also, the profiles observed in obese men closely resemble that observed in the oldest age group as presented in Figure 5.5. We confirmed differences across different BMI groups for average time spent in sedentary behaviours, following adjustment for age and sex. We found that moderately obese participants spent significantly more time in sedentary behaviours than normal-weight participants ($p<0.001$), and so did severely obese ($p<0.001$) and even overweight participants ($p<0.001$). We also found a strong significant difference between overweight and moderately obese participants

(p=0.0001); however, differences between normal-weight and underweight participants were not statistically significant (p=0.57).

Figure 5.7 Pitch (top panels), Roll (second row panels), Vector Magnitude High-Pass Filtered (third row panels), and sedentary time (bottom row panels) profiles (hourly averages) by time of the day in women and men, stratified by BMI categories (ranging from underweight [BMI: 16–18.5] to severely obese [BMI $\geq$35]).



5.4 Discussion

In this paper, we have explored the physical space in which physical activity occurs and described population differences in wrist movement and posture between men and women, age groups, BMI categories, and physical activity levels in a population sample of UK adults. Although higher activity was associated with lower pitch profiles, we observed the apparent paradox that older and more obese individuals who as groups are generally less active also spend more time at these postures, indicating that these groups either perform different types of activities or perform them at slower pace. Vector magnitude of movement intensity and pitch-roll angular features can all be considered direct measures of human behaviour, rather than estimates, as there is very little inference involved in deriving them; they have biomechanical meaning in their own right as also illustrated in Figure 5.8. The estimate of sedentary time, on the other hand, is not a direct measure but an estimate resulting from an inference but we have included it here to demonstrate the utility of combining directly measured features.

Remark: Including movement as well as pitch, roll (both indicating posture), and sedentary time estimates in our analysis allowed us to more comprehensively examine differences in human behaviour between time-of-day and weekdays and weekends, and illustrates the importance of taking all these features into consideration for large-population studies.

Non-surprisingly, our results suggest different wake-up times between weekdays and weekends; participants seem to wake up later during the weekends than weekdays. This information is of interest particularly given recent research suggesting that sleep irregularity may be a risk factor for cardio-metabolic disease [436]. The large differences in movement and postural measures between weekdays and weekends suggest differences in the type of activities that participants partake in between weekdays and weekends. These differences are particularly striking when comparing women and men. We found that women spend more time with their forearm elevated above horizontal than men do (53 % of their time vs. 36 % for men). Similarly, the pitch and roll profiles coincide with increases in movement around noon of the weekend days, pointing towards a behavioural pattern that could be suggestive of "weekend warrior" lifestyle, where participants tend to do most of their physical activity during the weekend. Further inspection of the data through visualization techniques (Figure 5.4) suggests that the activities participants engaged in strongly depended on time-of-day; it is apparent that the relative occupation of different physical spaces and the relationship between postures and movement changes drastically depending on the time of the day, indicative of engagement in different activity types.

We observed differences between men and women across most other substrata for both movement (vector magnitude) and posture (pitch and roll) measures, suggesting that men spent more time in postures that may be suggestive of sedentary behaviour than their female counterparts (sitting down, lying down). The inferred time estimate for sedentary behaviours (from vector magnitude and pitch), largely based on the methodology previously described by Rowlands et al [424], indicated that this was by far the most dominant behaviour across the whole population ($\sim$ 17 hours/day). However, younger individuals tended to spend less time than their older counterparts in these sedentary behaviours (suggesting more active lifestyles), and even starker differences were observed between different BMI groups; individuals with higher BMI spent the most time in sedentary behaviours, and we statistically confirmed that this was independent of age and sex. Movement and PAEE were both lower in the older age groups, a similar result to that observed in other population studies [437, 112, 162](2,33,34). We observed that older participants (60-65 age group) spend a large proportion of their time in postures that are similar to those with high BMIs, particularly in men. What was slightly paradoxical was that older and obese individuals spend more time at pitch angles generally associated with higher activity, ie with the arm below horizontal. As both movement and pitch are direct measurements of what the arm is physically doing, these results indicate true differences in activities, either as type or intensity or both. Using the sedentary time

estimation methodology, it was suggested that older and heavier individuals spent more time in sedentary behaviours. Future inference work on raw non-dominant wrist acceleration signals may further elucidate other differences, for example in the specific type of activity performed, including the separation of awake sedentary behaviour and sleep. Strengths of our study includes its standardised placement and 24-hour wear protocol which ensured greater certainty in the orientation of the accelerometer on each participant; that said, it is possible that some participants may have removed and replaced their device during the monitoring period. Still our results may provide guidance on probable axis orientation to other studies such as UK Biobank which do not have strict device orientation protocols. Another strength was that both wrist acceleration and PAEE was assessed simultaneously, thus providing more accurate stratification by PAEE levels; however a limitation of our work is that we only measured physical activity during one week of monitoring, and this may not be representative of habitual behaviour in this population. Another potential limitation is the separation between static and dynamic wrist acceleration; as has been previously addressed, the high- and low-pass filter parameters does not perfectly discriminate between static and dynamic and a small proportion of real movement will be missed during rapid rotations [438]. Nonetheless, this is likely to only bias the movement differences we observe towards the null, since younger and slimmer individuals are more able to produce more rapid movements, and it will likely not impact much on the postural measures, as the gravitational acceleration component is several orders of magnitude larger than residual movement in the low-pass filtered signal, thus still returning a valid estimate of the relative distribution of gravity in the three axes. Finally, in this work we describe hourly averages of behaviour and movement; however, there are alternatives to describing the underlying variability of these hourly-periods based on distributional statistics that were not explored in this paper.

5.4.1 Conclusions

In conclusion, we found that direct measures of accelerometry-derived forearm angles provide biomechanically relevant information alongside the more well-established movement intensity metrics such as vector magnitude to better characterize objectively measured physical activity in free-living conditions. Movement is more likely to occur, on average, at forearm angles below horizontal but despite older and heavier individuals moving less, these individuals still spend more time at lower forearm angles, suggesting population differences in style of movement which may be important for other health outcomes. These findings suggest that

these postural assessment methods for wrist accelerometers are valuable complements to traditional intensity based measures as explored in Chapter 1.

Figure 5.8 Raw triaxial wrist Acceleration, forearm Pitch and Roll Profiles (postures) for typical daily activities. From top to bottom: lying, walking, sitting and cycling.

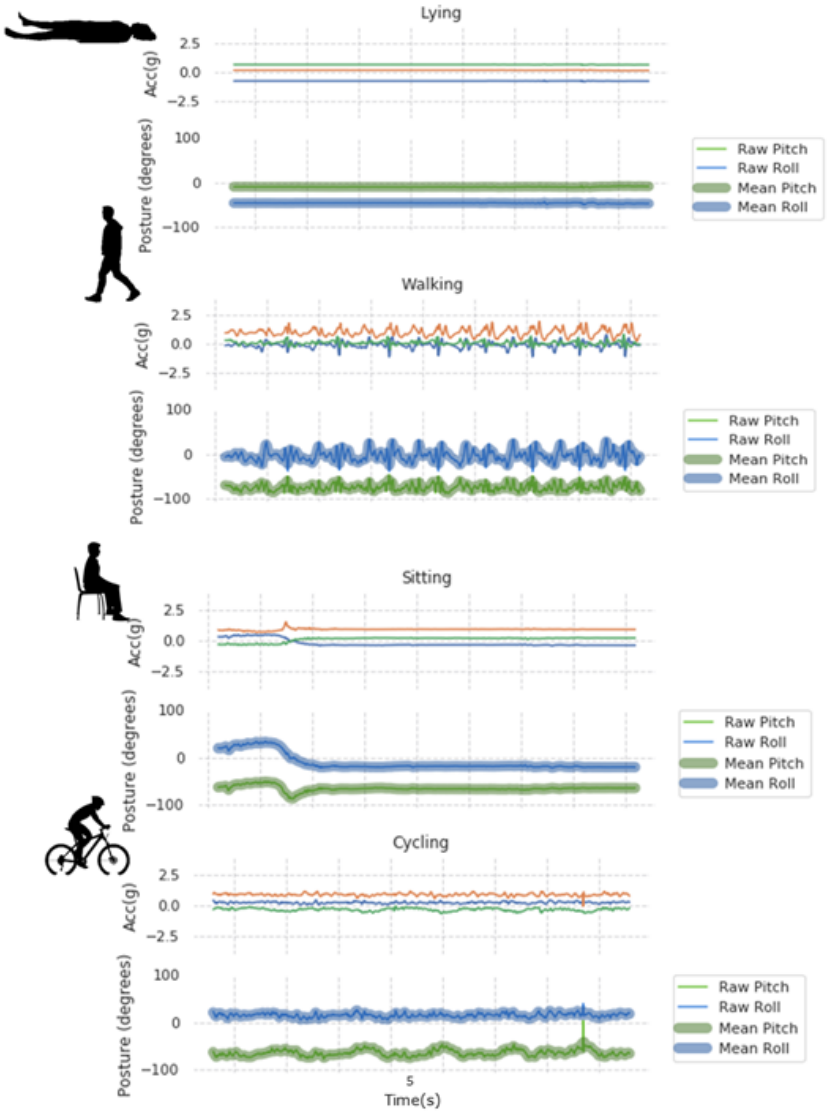
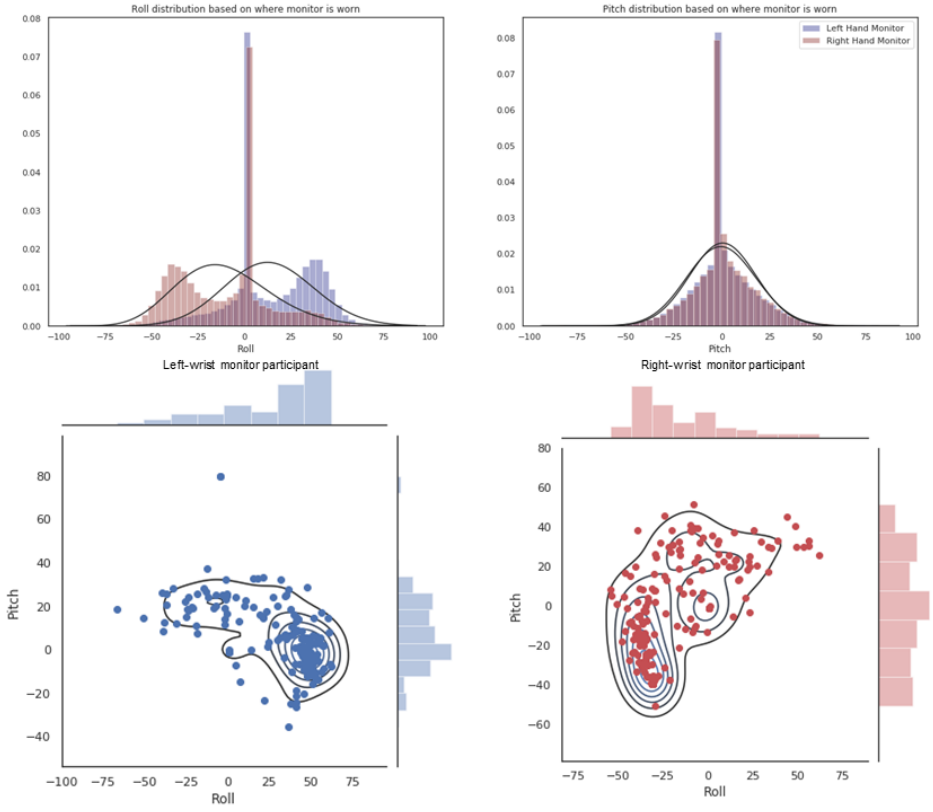


Figure 5.9 Untransformed Pitch and Roll distributions (full population), stratified by left versus right-hand accelerometer wear (top panel). The two plots underneath show examples of pitch-roll distributions from two participants wearing the accelerometer on their left (in blue) and right (in red) hand, respectively (each point is an hourly average).



CHAPTER 6

SELF-SUPERVISED TRANSFER LEARNING OF PHYSIOLOGICAL REPRESENTATIONS FROM LARGE SCALE FREE-LIVING WEARABLE DATA

Publications

This project is under submission for publication at the 35th AAAI Conference in Artificial Intelligence.

Contributions

I planned this project and devised the analysis plan in collaboration with Dimitris Spathis and our supervisors. I am truly appreciative of Dimitris' careful and thorough teaching on the use of custom deep learning methods and proper structuring of large analysis pipelines. Dimitris designed most of the deep learning experiments and taught me through this project how to deploy them at scale. My contributions to this work from an analysis standpoint focused on the data pre-processing and evaluation part of the project. We wrote the resulting manuscripts together which has been adapted for this chapter.

6.1 Summary

Background

Wearable devices such as smartwatches are becoming increasingly popular with consumers as well as a standard tool for large-scale public health studies. However, continuous and accurate monitoring of heart rate (HR) with these devices remains expensive and noisy in free-living conditions. Most wrist-worn wearable sensors are equipped with cheap accelerometers, which have proven to be reliable tools for physical activity monitoring. Further, current single modality sensors can tell us when and how much one particular behaviour, like exercise takes place as showcased in Chapter 5. State-of-the-art multimodal sensors coupled with deep learning could inform the effects that these behaviours have in the user's physiology, helping inform better decisions about our bodies and lifestyles.

Methods

In this work, we exploit the underlying physiological relationship between human activity and HR responses to forecast HR from accelerometry data. This approach offers an affordable, scalable and reliable way to monitor heart rate in free-living conditions without the need of an additional sensor. Similarly, labelling continuous accelerometer signals for use in supervised machine learning tasks can be prohibitively expensive and laborious. Therefore, we developed a *self-supervised* model that exploits HR data as a *supervisory signal* for activity data obtained from the accelerometer sensor. Through a deep neural network, we are able to forecast users' HR profiles using solely activity data as input. In addition, we propose a custom quantile loss function that accounts for the long-tailed HR distribution present on the general population given the variable fitness levels of individuals. Moreover, we conduct ablation studies to evaluate the impact of model components and input modalities across our experiments.

Results

We evaluate our model in the largest available free-living combined-sensing dataset (comprising $>$120k records of wrist accelerometer & wearable ECG). Our findings show that our model outperforms feature-based models and other baselines, for the personalized forecasting of heart rate. Additionally, we leverage the information captured through this self-supervised task by proposing a simple way to aggregate the learnt latent representations (embeddings) from the window-level to user-level. In doing so, we hypothesized that these embeddings would capture physiologically meaningful and personalized information given the nature of the self-supervised task. Notably, we show that the embeddings can generalize in various downstream tasks through transfer learning with linear classifiers. For example, they can be used to predict variables associated with individuals' health, fitness and demographic factors, outperforming simple bio-markers.

Conclusions

Overall, we propose the first multimodal self-supervised method for behavioral and physiological data. This method has implications for large-scale health and lifestyle monitoring and can be adapted to other parallel high-dimensional time-series tasks.

6.2 Background

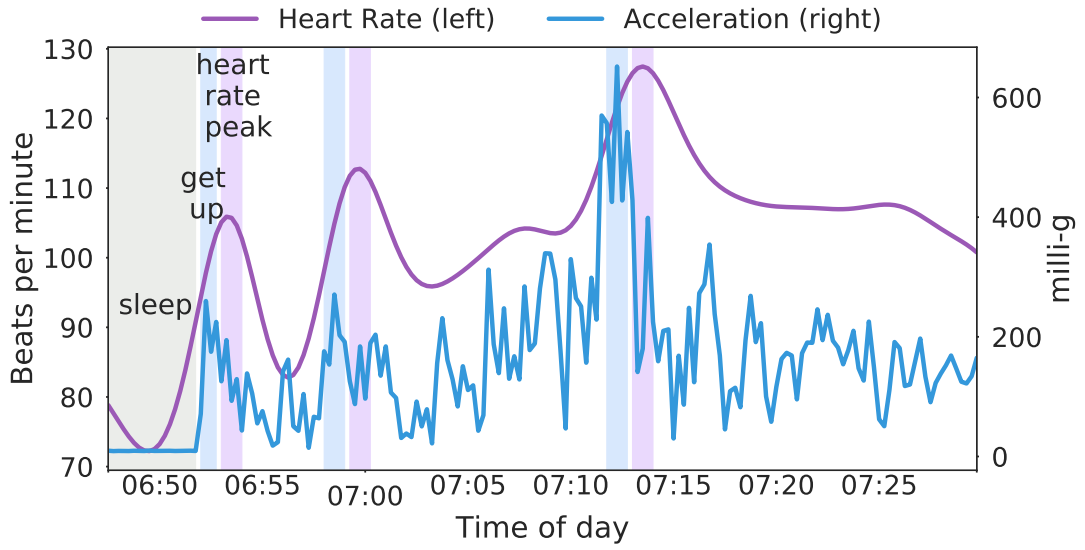
The advent of wearable technologies has given individuals and researchers the opportunity to ubiquitously and unobtrusively track everyday behavior. Thanks to the growth of internet-enabled wearable devices, sensor time series comprise a considerable amount of user-generated data [439]. However, extracting meaning from this data can be challenging, since sensors measure low-level signals (e.g., acceleration) as opposed to the more high-level events that are usually of interest (e.g., arrhythmia or obesity onset). Most wearable devices, particularly those that are wrist-worn, incorporate accelerometry sensors, which are valuable and very affordable tools to study physical activity patterns [397]. Indeed, the relationship between activity signals arising from wrist-worn accelerometers and physical activity has been studied in large population cohorts and is well established [112, 42, 419]. Physical behaviours are characterized by both movement and the associated cardiovascular response to movement (e.g. heart rate increases after exercise and the dynamics of this increase are dictated by fitness levels [440]) [1]. Heart Rate (HR) is a valuable marker not only for the study of specific physical behaviours, but also for the characterization and understanding of one's health and fitness level. In healthy individuals, HR responses to activity are defined by an increase in HR that is concurrent to the increasing intensity of the activity [441]. This relationship is conceptualized in Figure 6.1. HR responses to exercise have been shown to be strongly predictive of cardiovascular disease (CVD), coronary heart disease (CHD) and all-cause mortality [442]. However, due to limitations of sensing technologies, the relationship between continuous HR and health has not been thoroughly studied in free-living conditions.

A one-off measurement of resting HR (RHR) can be obtained using just one's fingers, either at the wrist or the side of the neck, or by using simple mobile applications requiring less than 4 minutes. By contrast, measuring continuous HR is expensive, often burdensome and noisy when engaging in higher levels of physical activity [443]. Photoplethysmograph (PPG) sensors, that infer HR from blood pulse waves, have been incorporated in smartwatches. However, signals obtained from PPG are corrupted by motion induced noise, which can impair their performance particularly under non-resting conditions [80]. Moreover, these sensors are still expensive, limiting their scalability and application in large-population studies.

In this work, we set to infer ECG-level quality HR signals by only utilizing accelerometry. Thereby, we provide an approach that *compresses* low-level sensor information while dealing with the idiosyncrasies inherent to this data, such as noise [444], the lack of annotations and labels associated to the raw signals [445], and long-tail distributions [446]. We present *Step2Heart*, a model that maps movement patterns from wrist accelerometery to HR responses, introducing the possibility of large scale, affordable and accurate inference of HR response in populations. We employ deep learning techniques that capture both the temporal dynamics

[1] Please note that throughout this work, we refer to activity, movement and acceleration interchangeably as signals obtained from wearable accelerometers.

Figure 6.1 Heart rate and acceleration temporal dynamics. Illustrative visualization of the relationship between movement and heart rate responses (randomly selected participant). Shaded areas show this lagging relationship.

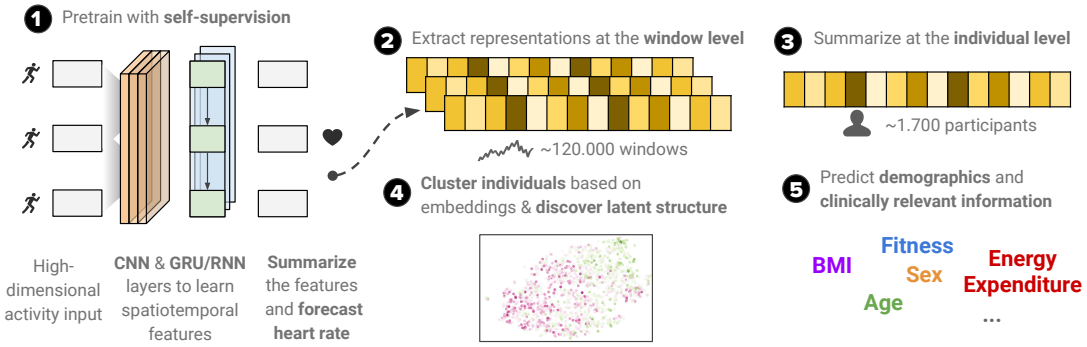


of sequential wearable sensor data and exploit the latent representations inherently present in this data. Recurrent and Convolutional neural networks have shown great promise in time-series modeling as they are able to flexibly capture non-linear relationships within sequential data [227], [447], [448], [449]. Importantly, these deep learning techniques do not require handcrafted feature engineering and are able to deal with noisy raw signals with diverse resolution such as those obtained from wearable sensors. These approaches have shown great promise in human activity recognition tasks using wearable sensor data [450–452].

Chapter Significance: In this chapter we devised a self-supervised framework that exploits the multimodal nature of modern wearable device data to generate personalized embeddings. In this work, we forecast heart rate resposes to movement as an auxiliary or pre-training tasks based on the hypothesis that through this task our model will learn valuable infomation regarding how each individual responds to movement or exercise. We then use the resulting embeddings in a number of downstream classification tasks to showcase the potential that these embeddings have when predicting health-related outcomes.

We also explore the latent information captured as a byproduct of the optimization of our architecture. Drawing parallels from the fields of natural language processing (NLP) and computer vision (CV), pioneers in representation learning [453], we posit that the behavioral and physiological signals captured by wearable sensors are appropriate and suitable for neural embeddings. In NLP and CV, researchers share pre-trained networks that can then be used to solve various downstream tasks. Inspired by the terminology used in [454], physiological signals display similar levels of *complexity* (it is not trivial to generate hand-crafted features) and *consistency* (movement is reflected as an increase in acceleration across all people) to

Figure 6.2 Schematic of model architecture and tasks.



NLP and CV. We believe that this could motivate a similar paradigm shift in the area of mobile health data especially given the privacy constraints associated with sharing such data. Instead, sharing models and embeddings would not directly expose participants' information and could accelerate research in a privacy-conscious way. Therefore, we position our model as a transfer learning model trained with self-supervision. *Self-supervised* learning is a training paradigm that exploits the intrinsic structure present in the data inputs, as explored in the introduction of this thesis. These models make use of large unlabelled data by learning objectives so as to get supervision from the data itself, using supervised loss functions [455]. Most importantly, the valuable intermediate representations capture intrinsic semantic meaning that can then be used for a variety of downstream tasks. A noteworthy multimodal application of this approach was learning to predict the audio of images and then using these representations to solve image recognition tasks [456]. In this work, we propose the first multimodal self-supervised model for behavioral and physiological data.

Remark: This chapter puts forward four key technical contributions:

- We propose a novel *self-supervised* approach to forecast HR responses from movement data using wrist-worn accelerometers. Through this architecture, our model learns *physiologically meaningful* user-level representations that can then be used for a variety of practical downstream tasks that are *personalized* to the users' unique physiology.
- We evaluate this model in what, to the best of our knowledge, is the largest dataset of its kind, including over 1,700 participants with combined wearable ECG and wrist accelerometry for a week. Our model outperforms a set of benchmarks and we perform ablation tests to show the performance of different input modalities to the architecture. The best HR forecasting models achieve an average error of $\sim$9 beats per minute in free living conditions.
- We introduce a joint *loss function* that acts as a regularizer to traditional MSE by using several quantiles of the predictive density of the model in order to capture the long-tails of HR data observed in the real world.

- We perform a set of downstream, transfer learning tasks by aggregating the window-level features to user-level ones and showcasing the value captured by the learned *embeddings* through strong performance at inferring physiologically meaningful variables. For example, our models achieve an AUC of 70 for BMI prediction and an AUC of 80 for Physical Activity Energy Expenditure.

We envision our work having applications in facilitating the comprehensive monitoring of cardiovascular health and fitness at scale. Further, our models could be used to correct faulty HR readings of noisy sensors such as PPGs and broadly to characterize the objectively measured physical behaviours in large population cohorts. Some of the downstream classification tasks highlight the potential of these techniques for the monitoring of important health information, which is usually costly or burdensome to obtain (such as fitness or obesity levels). The proposed model is summarized in Figure 6.2 and our code will be made publicly available.

6.3 Related work

Objective monitoring of physical behaviors. Large scale studies of physical activity leveraging mobile devices' built-in accelerometers have shown promise as global physical activity surveillance tools, demonstrating inequality across different countries and world regions [457]. Mobile and wearable sensors allow for continuous and ubiquitous monitoring of an individual's physical activity profiles, which combined with cardio-respiratory information, provides valuable insights into that individuals' health and fitness status [458]. Hence, the possibility of measuring individuals' physiological characteristics in free-living conditions is of great interest for research, clinical and commercial applications.

Machine learning for wearable sensing. Recently, advances in deep learning architectures for sequential modeling based upon wearable and mobile sensing have been used for health and fitness predictions and recommendations. For example, *FitRec*, an LSTM-based approach to modelling HR and activity data for personalized fitness recommendations was able to learn activity-specific contextual and personalized dynamics of individual user HR profiles during exercise segments [459]. This approach is helpful but requires prior segmentation of activities, which can be a constraint when applying these techniques in free-living, unconstrained conditions. Additionally, previous work has explored forecasting heart rate from movement data, however this was done on a much smaller scale (3 users) and used PPG sensors instead of the more accurate ECG as ground-truth [460].

Self-supervised pre-training. Recent work using self-supervised learning has shown state-of-the-art results in computer vision [65, 461], signal processing and natural language processing [455]. Use cases in wearable and mobile sensing have been limited to human activity

recognition using mobile devices [69] and emotion recognition using ECG data [462]. Our work is also inspired by the cardiovascular signature network introduced by Hallgrimmson and colleagues [463]. However, this is an auto-encoder based approach requiring a historical input of 1-month of data for its prediction which renders the whole setup not feasible for real time applications. Furthermore, the data used is much more aggregated than the data presented here. Overall, the generalizability of the learned embeddings is an under-explored area with some recent promising results in hospital operation room data [454], while attributes like gender and age have been proved to be predictable with wearable embeddings [464].

6.4 Methods

In this section, we provide with a brief introduction to the problem formulation and notation used and then explore the model architecture and the associated methods proposed in this chapter.

6.4.1 Problem formulation and notation

For this work, we assume N samples, an input sequence $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_N) \in \mathbb{R}^{N \times T \times F}$ and a target heart rate response $\mathbf{y} = (\mathbf{y}_1,...,\mathbf{y}_N) \in \mathbb{R}^N$. Additionally, we also consider contextual metadata like the hour of the day $\mathbf{M} = (\mathbf{m}_1,...,\mathbf{m}_N) \in \mathbb{R}^{N \times F}$. We use the same length T for all sequences in our model. However, this sequence length is not a requirement and can be adapted based on the requirements of the task at hand or the granularity of the data. The intermediate representations of the model after training are $\mathbf{E} = (\mathbf{e}_1,...,\mathbf{e}_N) \in \mathbb{R}^{N \times D}$ where D is the latent dimension. These embeddings are aggregated at the user level $\tilde{\mathbf{E}} = (\tilde{\mathbf{e}_1},...,\tilde{\mathbf{e}_N}) \in \mathbb{R}^{\frac{N}{U} \times D}$, where U is the number of users, in order to predict relevant outcome variables $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}_1},...,\tilde{\mathbf{y}_N}) \in \mathbb{R}^N$. Our full notation is summarized in Table 6.1. We employ two representation learning tasks: self-supervised pre-training and a downstream transfer learning task.

Pre-text task: self-supervised pre-training and HR forecasting. Given the accelerometer input sensor sequence $\mathbf{X}$ and associated metadata $\mathbf{M}$, predict the target HR $\mathbf{y}$ in the future. The input and target data shouldn't share temporal overlap in order to leverage the self-supervised paradigm and learn to predict the future. Motivated by population differences in heart rates, here we propose a custom *quantile regression loss* to account for the tails of the distribution. This task by itself can be used for a reliable and real-time estimation of HR based on activity data.

Downstream task: transfer learning of learned physiological representations. Given the internal representations $\mathbf{E}$ –usually at the penultimate layer of the aforementioned neural network [465]–, predict relevant variables $\tilde{\mathbf{y}}$ regarding the users' fitness and health using traditional classifiers (e.g. Logistic Regression). Inspired by the associations between word and

Notation	Description
$\mathscr{D}_{train}$, $\mathscr{D}_{test}$	training and testing set for the forecasting task
$\mathbf{X}$, $\in \mathbb{R}^{N \times T \times F}$	input sensor sequences
$\mathbf{M}$, $\in \mathbb{R}^{N \times F}$	input user metadata
$\mathbf{y}$, $\in \mathbb{R}^{N}$	target heart rate response
N	number of data points (samples)
T	length of input sequence
F	number of features (attributes)
U	number of users
$\tilde{\mathscr{D}}_{train}$, $\tilde{\mathscr{D}}_{test}$	training and testing set for the transfer learning task
θ	parameters (weights) of a trained neural network
D	dimension of latent space embedding
$\mathbf{E}$, $\in \mathbb{R}^{N \times D}$	embeddings matrix learned from activity to heart rate mapping
$\tilde{\mathbf{E}}$, $\in \mathbb{R}^{\frac{N}{U} \times D}$	embeddings matrix learned like $\mathbf{E}$ (aggregated at the user level)
$\tilde{\mathbf{y}}$, $\in \mathbb{R}^{\frac{N}{U}}$	target variable for transfer learning (user level)

Table 6.1 **Notation.**

document vectors in NLP [466], we develop a simple aggregation method of sensor windows to the user level. This is a common issue in the literature [467].

6.4.2 Model architecture

As shown in Figure 6.2 we propose *Step2Heart*, a deep neural network for HR forecasting and transfer learning. Its layers receive high-dimensional activity inputs along with associated metadata and learn spatio-temporal dynamics in order to accurately predict HR responses. It uses stacked convolutional (CNN) and recurrent (RNN) layers building upon architectures like *DeepSense* [468], which have been proven state of art in mobile sensing. Here we present each component of the model. An overview of the overall method is given as a pseudocode in Algorithm 1.

6.4.2.1 CNNs to learn spatial features

Given an input sequence $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_N)$, it passes through a stack of CNN layers that scan over the sequences with 1D windows and learn filters $f : \{0, ..., k-1\} \in \mathbb{R}$. The convolution operation F of a sequence element s is defined as

$$F(s) = (\mathbf{x} * f)(s) = \sum_{i=0}^{k-1} f(i) \cdot \mathbf{x}_{s-i} \tag{6.1}$$

where k is the filter size, $s - i$ records the convolution step and $*$ denotes the convolution operator. Please note that the 1D window learns patterns across all the parallel features of the 3D input tensor $\mathbf{X}$.

6.4.2.2 RNNs to learn temporal features

The learned filters of the CNNs are then fed into stacked RNNs. Specifically we employ a fast variant of RNNs known as Gated Recurrent Units (GRU) [447]. The GRU has a reset gate r and an update gate z which change the hidden state h at each time step. The update functions are as follows:

$$\begin{aligned} \mathbf{r}_t &= \sigma(W_r \mathbf{x}_t + U_r \mathbf{h}_{t-1} + \mathbf{b}_r) \\ \mathbf{z}_t &= \sigma(W_z \mathbf{x}_t + U_z \mathbf{h}_{t-1} + \mathbf{b}_z) \\ \tilde{\mathbf{h}_t} &= tanh(W \mathbf{x}_t + U(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}) \\ \mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}_t} \end{aligned} \tag{6.2}$$

where matrices W, U and $\mathbf{b}$ are model parameters and biases respectively, σ is a sigmoid function, and $\odot$ element-wise multiplication. The stacked GRUs output sequences which correspond to latent temporal features.

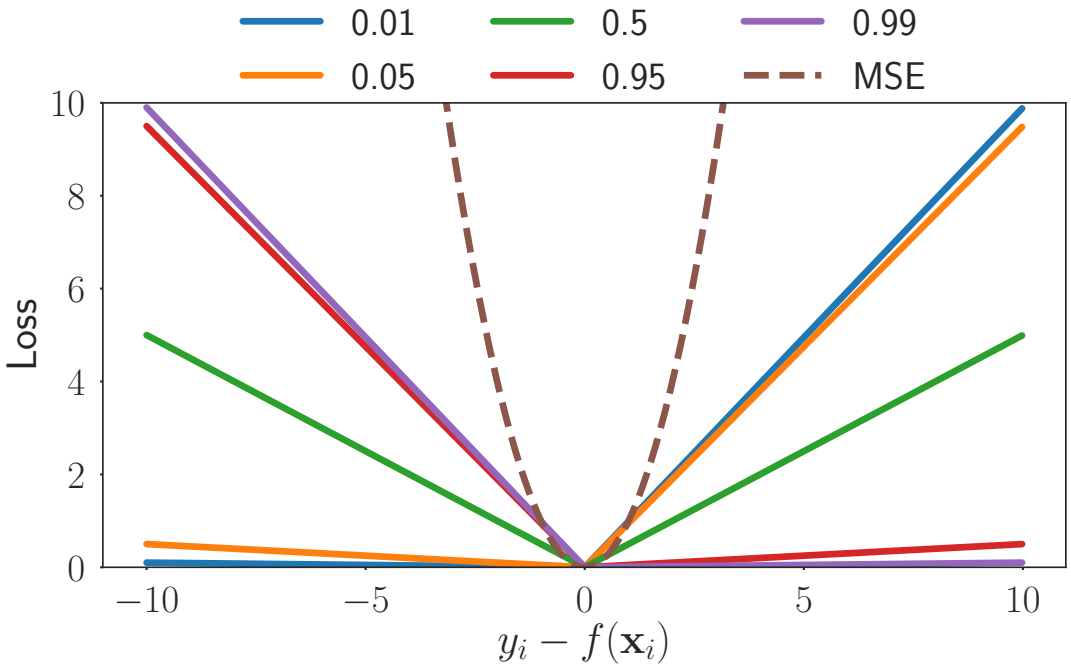
Algorithm 1: *Step2Heart* model pseudocode

Input : $\mathbf{X}$ (sensors), $\mathbf{M}$ (metadata), $\mathbf{y}$ (target HR)
Output : $\tilde{\mathbf{E}}$ (user-level embedding), $\tilde{\mathbf{y}}$ (target variable)
while *neural network θ not converged* **do**
    pass $\mathbf{X}$ through CNN/RNN layers (eq. 6.1 & 6.2);
    pass $\mathbf{M}$ through reLU layers;
    concatenate outputs in $\mathbf{E}$;
    forecast & backpropagate with joint loss $\mathcal{L}$ (eq. 6.5);
end
use trained network θ to extract embeddings $\mathbf{E}$;
aggregate $\mathbf{E}$ to the user-level $\tilde{\mathbf{E}}$ with average pooling;
train a linear model to predict target variables $\tilde{\mathbf{y}}$;

6.4.2.3 Pooling and prediction

Then, the GRU output $\mathbf{h}_t$ passes through a pooling layer that performs global element-wise averaging in order to summarize all the timesteps of the 3D tensor to a 2D matrix. If needed, the representation after the pooling operation can be concatenated with other features or metadata after passing through feed forward *ReLu* layers. We also refer to this representation at the

Figure 6.3 Quantile vs MSE loss. Illustration of the relationship between the prediction and the loss with respect to the shapes of the MSE and various levels α of quantiles. Simulated data, the true value is $y_i = 0$.



penultimate layer, **E**, or *embeddings* matrix. Lastly, the final layer is a feed forward neural network with a linear activation which is appropriate for regression tasks.

6.4.2.4 Custom loss function

Heart rates vary across large populations. As such, some individuals may reach very low (<50 bpm, at rest/sleeping) or high (>180 bpm during vigorous exercise) [469] generating very long tails on the heart rate distribution. In traditional regression, the aim is to minimize the squared-error loss function or MSE $\mathcal{L}_{MSE}(\mathbf{y}, \mathbf{f}) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y_i - f(\mathbf{x}_i))^2$ to predict a single point estimate, similarly, quantile regressions aim to minimize the quantile loss in predicting a certain quantile. As such, the higher the quantile, the more the quantile loss function penalizes underestimates and the less it penalizes over estimates.

The loss for an individual data point in quantile regression is defined by:

$$\mathcal{L}(\xi_i|\alpha) = \begin{cases} \alpha\xi_i & \text{if } \xi_i \geq 0, \\ (\alpha - 1)\xi_i & \text{if } \xi_i < 0. \end{cases} \tag{6.3}$$

where α is the required quantile (between 0 and 1) and

$\xi_i = y_i - f(x_i)$, where $f(x)$ is the predicted (quantile) model and y is defined by the observed value for input x. A more compact version of Eq. (6.3) can be formulated as $\mathcal{L}(\xi_i|\alpha) = max(\alpha\xi_i, (\alpha-1)\xi_i)$ where $\xi \in \mathbb{R}$ is the residual. As such, the average quantile loss over the whole dataset is:

$$\mathcal{L}_Q(\mathbf{y}, \mathbf{f}|\alpha) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y_i - f(\mathbf{x}_i)|\alpha) \tag{6.4}$$

The quantile loss (or tilted/pinball loss in the literature) can be seen as *tilted* version of the l_1 loss which estimates the unconditional median. Instead, if a prediction falls below a given quantile (e.g. $\alpha = 0.10$), the residual is scaled (or tilted) by its probability α. Thus, we can obtain the conditional quantile by minimizing the empirical $\mathcal{L}_Q$ loss. This formulation is inspired by similar loss functions applied to transportation problems [470] as well as reinforcement learning [471].

In practice, we are interested in different quantile levels for the predicted probability distribution, not only one. Let $\{a\}_{j=1}^{J}$ be a set of J quantiles (e.g. 0.05, 0.10, ..) we propose a joint loss function that leverages the $\mathcal{L}_{MSE}$ and $\mathcal{L}_Q$ loss for an arbitrary number of quantiles:

$$\mathcal{L}_{MSE+Q} = \frac{1}{N}\sum_{i=1}^{N}\left((y_i - f(\mathbf{x}_i))^2 + \sum_{j=1}^{J} max\Big(\alpha_j(y_i - f(\mathbf{x}_i)^{(\alpha_j)}), (\alpha_j - 1)(y_i - f(\mathbf{x}_i)^{(\alpha_j)})\Big)\right) \tag{6.5}$$

which can be seen as a sum of the MSE and the respective quantile losses, represented in one scalar. This scalar is used as the new backpropagation objective.

In Figure 6.3 we use a toy example to illustrate the differences between the MSE and Quantile loss: the former increases very fast in case of outliers, whereas the latter is more robust. For the individual quantiles, we observe that for very extreme values (e.g. 0.01 or 0.99) the loss skews significantly assigning high penalties to underestimation and overestimation, respectively. In our context, very athletic or sedentary people can be considered as long-tail outliers and we want our models to account for it. Intuitively, the proposed loss can be seen as a combination of multiple objective functions where the second term acts as a regularizer for the MSE. During our experiments in next sections we apply different ablations of these terms to evaluate their impact.

Feature	Seq.	Inp.	Unit
Sensor			
Acceleration	✓	✓	m/s^2
Heart Rate	✓	✗	Beats per Minute (BPM)
Timestamp	✓	✓	N/A
Metadata			
UserID	✗	✗	N/A
Height	✗	✗	Meters
Weight	✗	✗	Kilograms
Sex	✗	✗	Male–Female
Resting HR	✗	✗	BPM
VO_2max	✗	✗	mL/min · kg
Derived			
Triaxial Acceleration	✓	✓	m/s^2
ENMO	✓	✓	milli-g
VM-HPF	✓	✓	milli-g
PAEE	✗	✗	$J/min\cdot$ kg
Body Mass Index (BMI)	✗	✗	kg/m^2
Month, Hour	✓	✓	[-1,1] cos/sin transform

Table 6.2 **Data description**. *Seq.* denotes sequential measurements (timeseries), while *Inp.* the inputs to the forecasting model. *Triaxial Acceleration*: `mean`, `std` of x, y, z axes, *ENMO*: `mean` of Euclidean Norm Minus One. *VM-HPF*: `mean`, `min`, and `max` of Vector Magnitude High-Pass Filter. *PAEE*: Physical Activity Energy Expenditure.

6.5 Evaluation

6.5.1 Dataset

Participants were recruited from general practice lists in and around Cambridgeshire in the East of England to the Fenland Study, a population-based study designed to investigate interactions between environmental and genetic factors in determining obesity, type 2 diabetes, and related metabolic disorders in younger and middle-aged adults between the ages of 35-65 [396]. Exclusion criteria included prevalent diabetes, pregnancy or lactation, inability to walk unaided, psychosis or terminal illness (life expectancy of ≤ 1 year at the time of recruitment).

After a baseline clinic visit, a subsample of 2,100 participants were asked to wear a combined heart rate and movement chest sensor (Actiheart, CamNtech, Cambridgeshire, UK) and a wrist accelerometer (GeneActiv, ActivInsights, Cambridgeshire, UK) on their non-dominant wrist.

Before the free-living monitoring period, participants performed a treadmill test to establish their individual heart rate response to a submaximal test. These measurements were used to produced calibration parameters to inform a branched equation model of PAEE, which has been validated against instantaneous PAEE (intensity) from indirect calorimetry. Weight was measured to the nearest 100g using a calibrated scale (TANITA model BC-418MA; Tanita, Tokyo, Japan) and height was measured to the nearest 0.1cm using a calibrated wall-mounted stadiometer (SECA 240; Seca, Birmingham, UK). Body Mass Index (BMI) was calculated in kg/m^2

All participants provided written informed consent and the study was approved by the University of Cambridge, NRES Committee - East of England Cambridge Central). All experiments and data collected were done in accordance with the declaration of Helsinki.

6.5.1.1 Study protocol

The *mobile ECG* measured heart rate and uniaxial acceleration in 15-second intervals while the wrist device recorded 60 Hz triaxial acceleration. Participants were told to wear both monitors continuously for 6 complete days and were advised that both monitors were waterproof and could be worn during showering, sleeping or exercising. During a lab visit, all participants performed a treadmill test that was used to establish their individual response to a *submaximal test*, informing their VO_2max (maximum rate of oxygen consumption measured during incremental exercise). RHR was measured with the participant in a supine position using the *mobile ECG*. HR was recorded for 15 minutes and RHR was calculated as the mean heart rate measured during the last 3 minutes. These measurements were then used to calculate the Physical Activity Energy Expenditure (PAEE) [472].

	MSE	RMSE	MAE
Step2Heart (A)	144.61 (0.62)	12.02 (0.02)	9.23 (0.03)
Step2Heart (A/T)	143.65 (0.28)	11.98 (0.01)	9.21 (0.03)
Step2Heart (A/R)	91.76 (0.12)	9.57 (0.00)	6.92 (0.03)
Step2Heart (A/R/T)	**91.11** (0.37)	**9.54** (0.01)	**6.88** (0.02)
Baselines			
Global mean	250.99	15.84	12.46
User mean	186.05	13.64	10.40
XGBoost (A)	162.92 (0.20)	12.76 (0.00)	9.83 (0.00)

Table 6.3 **Forecasting task results.** Ablation test to compare the HR forecasting error using different input modalities and baselines. To make for a fair comparison please note that only the MSE loss ($\mathcal{L}_{MSE}$) is used as an objective for our models here. (A=acceleration, T=temporal features, R=Resting Heart Rate)

6.5.1.2 Pre-processing

All participant heart rate data collected during free-living conditions underwent pre-processing for noise removal. Similarly, all accelerometer data was auto-calibrated to local gravity, non-wear time was inferred and participants with less than 72 hours of wear were removed. Magnitude of acceleration was calculated through *high-passed filtered vector magnitude (VM HPF)* (expressed in milli-g/mg per sample). Both the accelerometry and ECG signals were summarized to a common time resolution of one observation per 15 seconds and no further processing to the original signals was applied.

Since what time it is can have a big impact on physical activity be it sleeping, commuting or even the season of the year, we encoded the sensor timestamps using *cyclical temporal features* T_f [473]. Here we encoded the month of the year and the hour of the day as (x, y) coordinates on a circle:

$$T_{f1} = sin\Big(\frac{2 * \pi * t}{max(t)}\Big) \qquad (6.6) \qquad\qquad T_{f2} = cos\Big(\frac{2 * \pi * t}{max(t)}\Big) \qquad (6.7)$$

where t is the relevant temporal feature (hour or month). The intuition behind this encoding is that the model will "see" that e.g. 23:59 and 00:01 are 2 minutes apart (not 24 hours). We perform ablation tests with and without these features during our experiments to evaluate their impact.

6.5.2 Training procedure

To create appropriate training batches for deep learning, we segmented the signals into fixed *non-overlapping* windows of 512 timesteps, each one comprising 15-seconds and therefore

yielding a window size of approximately 2 hours. We divided our dataset into training and test sets randomly using an 80-20% split with the training set then being further split into training and validation sets (90-10%). We ensured that the test and train set had disjoint user groups. Further, we normalized the data by performing min-max scaling on all features described on Table 6.2 (sequence-wise for timeseries and column-wise for tabular ones) on the training set and applying it to the test set. During training, the target data (HR bpm) is not scaled and the forecast is 15" in the future after the last activity input.

The neural network was built through a stack of 2 CNN layers of 128 filters each, followed by 2 Bidirectional GRU stacked layers of 128 units each (resulting in 256 features due to bidirectional passes). When using extra inputs (RHR or timestamp derived features), a *ReLu* MLP of dimensionality 128 was employed for each one and its outputs were concatenated with the GRU output. We trained using the Adam [474] optimizer for 300 epochs or until the validation loss stopped improving for 5 consecutive epochs [2]. The quantiles we used were [0.01, 0.05, 0.5, 0.95, 0.99] so that they equally cover extreme and central tendencies of the heart rate distribution. The XGBoost baseline's hyperparameters were found through 5-fold cross validation and were then applied to the test set. Likewise, in the transfer learning task, we followed the same procedure for Logistic Regression.

For the transfer learning task, we studied if the learned embeddings $\mathbf{E}$ can predict user variables ranging from demographics to fitness and health. Since a slightly lower number of users (1506) had sufficient fitness data obtained from the lab test visit, we report only their results (the users remained in the same train/test splits $\tilde{\mathscr{D}}_{train}$ / $\tilde{\mathscr{D}}_{test}$ as earlier). To create binary labels we calculated the 50% percentile in each variable's distribution for the training set and assigned equally sized positive-negative classes. The window-level embeddings were averaged with an element-wise mean pooling to produce user-level embeddings[3]. Then, to reduce overfitting, Principal Component Analysis (PCA) was performed on the training embeddings after standard scaling and the resulting projection was applied to the test set. We examined various cutoffs of explained variance for PCA, ranging from 90% to 99.9%. Intuitively, lower explained variance retained fewer components; in practice the number of components ranged from 10 to 160.

6.5.3 Baselines and metrics

For our baselines, we used naive lower bounds that require no models as well as modern ML regressors:

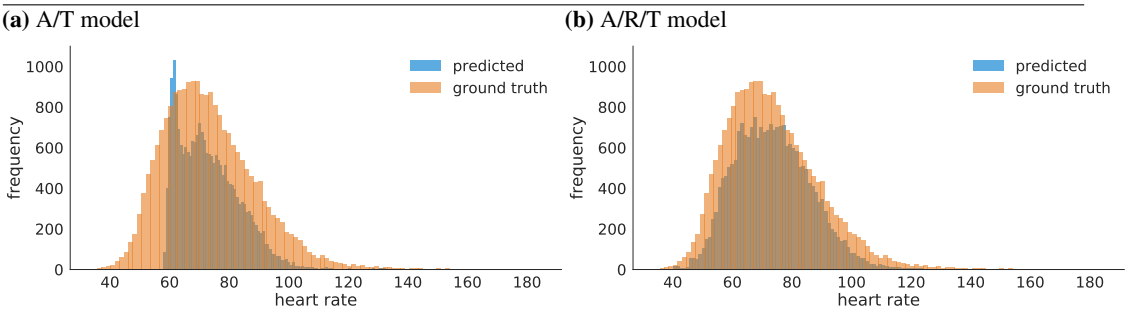
[2] hyper-parameter search was conducted with different layer numbers, unit sizes, learning rates and optimizers and we evaluated their impact on the validation set.

[3] we experimented with min, max and median pooling over embeddings but yielded consistently worse results across all variables.

	MSE	RMSE	MAE
Step2Heart (A/R/T)			
$\mathcal{L}_{MSE}$	91.11 (0.37)	9.54 (0.01)	6.88 (0.02)
$\mathcal{L}_{MSE+Q}$	90.94 (1.12)	9.53 (0.05)	6.90 (0.10)
$\mathcal{L}_{0.5*MSE+Q}$	**90.27** (0.53)	**9.50** (0.02)	6.81 (0.05)
$\mathcal{L}_{Q}$	92.0 (0.16)	9.59 (0.00)	**6.75** (0.02)

Table 6.4 **Loss function results.** Ablation test to compare the best performing model in terms of modalities (*Step2Heart* (A/R/T)) in regards to different loss functions.

Figure 6.4 Forecasting predictions (a-b). Prediction distributions (test set) using *Step2Heart*, showing the impact of including the (static) resting heart rate as input.



- **Global mean:** Predicts y_i at each time step as the global HR mean of the training set. This is a naive baseline that assumes all users have the same HR anytime but provides a good lower bound for this longitudinal dataset.

- **User mean:** *Personalized* baseline obtained by predicting y_i at each time step as the mean value for all the user's $\mathbf{X}$ in the training set. This is similar to the previous baseline but considers the entire heart rate range of each user over the study week.

- **Autoencoder:** A convolutional autoencoder which compresses the input data $\mathbf{X}$ with a reconstruction loss. This baseline uses movement data only and the intuition behind this is to assess whether *Step2Heart* learns better representations due to learning to map movement to heart rate $\mathbf{X} \rightarrow \mathbf{y}$. To make a fair comparison, it has similar number of parameters to the self-supervised models and we use the bottleneck layer to extract embeddings (128 dimensions). This baseline is used only for the transfer learning experiments.

- **Gradient Boosting (XGboost)** (A): gradient boosting machines are among the best performing ML methods [475]. We use it as a comparison to demonstrate that the temporal modeling capabilities of our architecture through CNN/GRUs outperform the best ML baselines. Since XGboost cannot work directly on timeseries, we extracted the following statistical features from the sensor windows: mean, std, max, min, percentiles (25%, 50%, 75%) and the slope of a linear regression fit. The final feature vector consists of 80 features.

Outcome	AUC											
	Autoencoder				*Step2Heart*$_{A/T}$				*Step2Heart*$_{A/R/T}$			
PCA*	90%	95%	99%	99.9%	90%	95%	99%	99.9%	90%	95%	99%	99.9%
VO_2max	52.6	52.6	59.6	61.8	58.6	60	63.9	64.5	**68.3**	67.8	68	68.2
PAEE	69.6	70.0	70.2	71.8	74.7	74.7	77.5	76.8	78.2	79.2	**80.6**	79.7
Height	60.8	60.3	75.9	79.4	66	67.4	77.4	**82.1**	70.3	74	80.5	81.3
Weight	56.5	56.2	70.3	72.1	65.7	67.6	75	77.2	69.9	70.7	**77.4**	76.9
Sex	66.7	67.0	86.5	89.7	72.3	72.9	87.1	93.2	76.2	81.5	91.1	**93.4**
Age	46.2	46.3	53.9	59.5	55.0	61.7	66.2	66.9	61.1	63.8	67.3	**67.6**
BMI	51.6	51.5	60.1	61.2	62.8	63	68.2	67.6	64.7	66.1	67.8	**69.4**
Resting HR	49.1	49.4	55.8	55.4	56.7	56.6	**62.7**	61.7	N/A			

Table 6.5 **Transfer learning results**. Performance of embeddings in predicting variables related to health, fitness and demographic factors. A random baseline yields an AUC of 50. (*percentage of explained variance by compressing the dimensionality of embeddings with PCA)

Given the sequential nature of the regression forecasting task, we use standard metrics such as the Root Mean Squared Error (RMSE), Mean Squared Error (MSE) and Mean Absolute Error (MAE) for our evaluation. While MSE–RMSE put importance on the biggest errors, MAE gives the same importance to each error and is more robust to outliers. For the transfer learning task, the evaluation metric is the Area under curve (AUC). These metrics can described as follows:

$$RMSE = \sqrt{\frac{1}{|N_{test}|} \sum_{y \in \mathscr{D}_{test}} \sum_{t=1}^{N} (y_t - \hat{y}_t)^2} \tag{6.8}$$

$$MSE = \frac{1}{|N_{test}|} \sum_{y \in \mathscr{D}_{test}} \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 \tag{6.9}$$

$$MAE = \frac{1}{|N_{test}|} \sum_{y \in \mathscr{D}_{test}} \sum_{t=1}^{N} |y_t - \hat{y}_t| \tag{6.10}$$

6.6 Results

6.6.1 Forecasting

We consider different ablation tests for *Step2Heart* as well as several baselines and report the average and standard deviation of 3 runs. All results are evaluated on the test set. For our

ablation tests we consider the same model with different inputs: acceleration features only (A), with temporal features (A/T), with resting heart rate (A/R) and with both temporal features and resting heart rates (A/R/T).

Impact of the Resting Heart Rate. All results are summarized in Table 6.3. *Step2Heart* outperforms all baselines for this forecasting task and, when including temporal features and resting heart rate (*Step2Heart*(A/R/T)), all performance metrics improve, resulting in an RMSE of 9.54. We note that the RMSE is probably the most interpretable metric since it directly translates to the error in HR beats per minute. Figures 6.3a and 6.3b depict the improvement in performance when including the resting heart rate and temporal features. Namely, while the (A/T) model seems to struggle with the prediction of HR which is below 60 BPM (presumably fit individuals or during sleep time) the (A/R/T) model manages to capture this tail and as well approximate better the long tail of the high BPM. Therefore, given the acceleration input, the addition of the RHR appears to be the most significant input, improving the RMSE by ~ 2.5 and validating previous research that highlights RHR as a powerful bio-marker [476]. Regarding the temporal features, their contribution is marginal when combined with an acceleration model (A/T), but they achieve the lowest error in combination with the RHR (A/R/T), showcasing the impact of the seasonality and circadian rhythms.

Implicit personalization. Interestingly, the baselines also serve to reinforce the importance of personalized approaches for this type of tasks as the user mean baseline vastly outperforms the global mean. Also our models implicitly learn personalized patterns outperforming both the XGboost model and model-free baselines. Given the strong results of the embeddings in demographic prediction we present in the next section, we postulate that these models learn personalized features that would not be possible with other methods that –for example– require user-specific layers and might not scale in large-scale datasets [477].

Impact of the joint loss. When comparing different loss functions with the best performing model *Step2Heart*(A/R/T), we see (Table 6.4) that the proposed loss function better captures the long tails of HR. The lowest error, 9.5 RMSE, is achieved when weighting the MSE loss with the rest of the quantiles ($\mathcal{L}_{0.5*MSE+Q}$). Notably the pure quantile model achieves the best MAE of 6.75. We understand that a model optimized with the MSE loss would achieve better MSE score and a model including the 50% quantile would optimize the MAE score. Thus, for this experiment we evaluate the impact of the losses *across* all 3 metrics. In this case, the joint losses achieve the best results; the $\mathcal{L}_Q$ model may achieve the best MAE but predicably falls short in the other metrics. Given the overlapping standard deviations of the joint models ($\mathcal{L}_{0.5*MSE+Q}$ and $\mathcal{L}_{MSE+Q}$) we consider both to be our best models, however we pick the former as the one with the lowest average error.

Practical applications. This approach may be employed to estimate HR in consumer devices and large population studies that collected accelerometry data. Moreover, it could be used for the systematic minimization of motion-induced HR measurement error in wearable devices that

integrate optical heart rate through PPG [80] and movement sensing or for more energy-efficient wearable sensing deployments [460].

6.6.2 Transfer learning

For this set of results, we used the model with the best performance as shown above ($\mathcal{L}_{0.5*MSE+Q}$), extracted the embeddings as described in sec. 6.5.2 and trained linear classifiers (logistic regression) for different outcomes using the same embeddings.

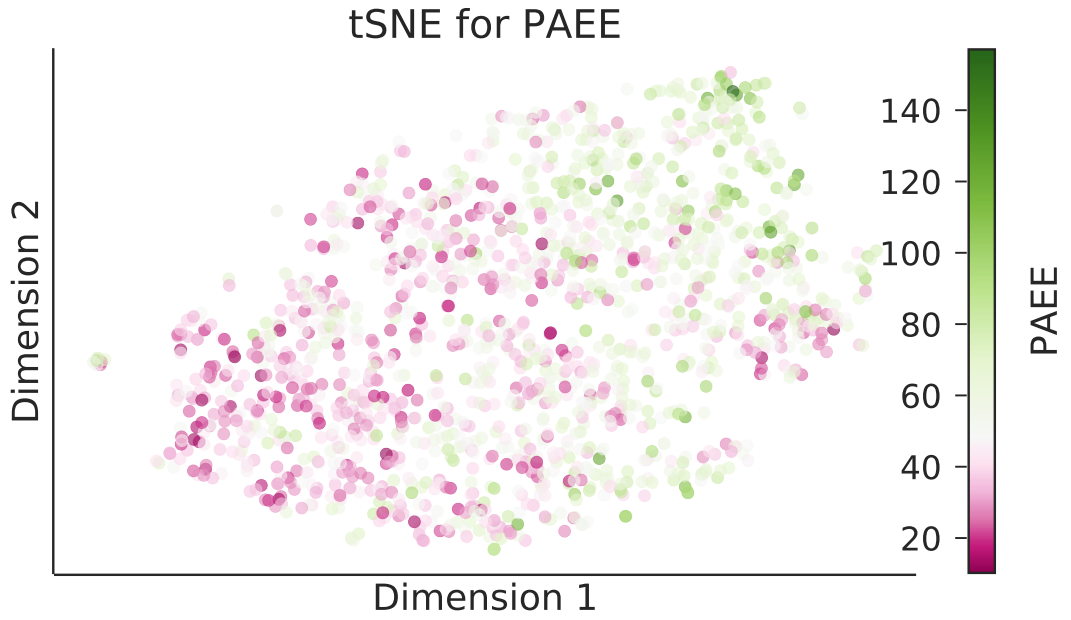
Generalizing in downstream tasks. Quantitatively, the embeddings achieved strong results in predicting variables like user's sex, height, PAEE and weight (93, 82, 80 and 77 AUC respectively). Also, BMI, VO_2max and age are moderately predictable (70 AUC). The pure acceleration model (A/T) moderately predicts Resting HR (62 AUC), but this does not apply to the (A/R/T) since it already includes the RHR as input. Generally, the A/R/T model outperforms the A/T model showing that using the RHR as input is helpful, as discussed in the previous sections.

Our results validate previous studies like [463] with different data, which reported achieving 70 AUC in BMI and age prediction by auto-encoding wearable data. As a simple baseline, we followed their idea of using the RHR as a single predictor and we could not surpass an AUC of 55 for BMI and age. Also, the autoencoder baseline, which learns to compress the activity data, under-performs when compared to *Step2Heart*$_{A/T}$, illustrating that the proposed task of mapping activity to HR captures the physiological state of the user, which translates to more generalizable embeddings. We note that both approaches operate only on activity data on inference time. This shows that the embeddings carry richer information than single biomarkers or modalities by leveraging the relationship between physical activity and heart rate responses. Obtaining these variables in large populations can be valuable for downstream health-related inferences which would normally be costly and burdensome (for example a VO_2max test requires expensive laboratory treadmill equipment and respiration instruments). Additionally, PAEE has been strongly associated with lower risk of mortality in healthy older adults [478]. Similarly, VO_2max is prospectively associated with the incidence of type 2 diabetes [88].

Implicit personalization. Interestingly, the baselines also reinforce the importance of personalized approaches as the user mean baseline vastly outperforms the global mean. Our models implicitly learn personalized patterns outperforming all baselines. Given the strong results of the embeddings in demographic prediction we present in the next section, we postulate that these models learn personalized features which would not be possible with other methods that –for example– require user-specific layers and might not scale in large-scale datasets [477].

Impact of the latent dimensionality size. From the representation learning perspective, we observe considerable gain in accuracy in some variables when retaining more dimensions

Figure 6.5 Model embeddings for transfer learning visualized with t-SNE. 2D representation of the embeddings for PAEE prediction. Color coding shows the extreme expenditures, since the median participant had PAEE 48 (white color). See Table 6.5 for full results.



(PCA components). More specifically, Sex and Height improve in absolute around +20 in AUC. However, this behavior is not evident in other variables such as PAEE and VO_2max, which seem robust to any dimensionality reduction. This means that the demographic variables leverage a bigger dimensional spectrum of latent features than the fitness variables which can be predicted with a subsample of the features. These findings could have great implications when deploying these models in mobile devices and deciding on model compression or distillation approaches [479].

Visualizing the latent space. Qualitatively, we visualized the resulting *latent* space in 2D with t-Distributed Stochastic Neighbor Embedding (t-SNE) [480] as shown in Figure 6.5. In this setup, we used both the embeddings of the train and the test sets together since we want to illustrate the structure of the entire dataset. We found that many of the inferred variables, like the depicted PAEE cluster in their own specific regions. We decided to color code and focus on the extreme PAEE users in order to illustrate the fact that most normal users are spread in the middle but high/low PAEEs are diametrically opposed. This structure and separability is hypothesized to be due to the nature of the embeddings learned through the self-supervision process. These visualizations can help us understand common behaviours (similar users are neighbors in the latent space), would allow for risk stratification and potentially suggest interventions to specific groups (e.g. nutrition or exercise advice to high-risk obesity onset cluster).

6.6.3 Discussion

Our results showcase the generalizability of the proposed models to solve different tasks relevant to physiological and behavioral data. Our models achieve competitive performance in real-time HR forecasting as well as in more abstract tasks such as cardio-respiratory fitness prediction and demographics inference.

With regards to the HR forecasting, we consider the error acceptable for real-world deployments, especially in cases of energy-constraint environments where the heart rate sensors could be prohibiting (an accelerometer consumes substantially less power). In our future work, we will assess the feasibility of the deployment of such model and examine its performance in different conditions (e.g. HR is generally steadier during sleep and that may affect the average error).

Perhaps more interestingly, are the transfer learning results, as they showcase the potential that this method holds for the generation of participant-specific representations. Through self-supervised learning, we can leverage this unlabelled data to learn meaningful representations that can generalize in situations where ground truth is inadequate or simply infeasible to collect due to high costs. Such scenarios are of great importance in population health where we may be able to achieve clinical-grade health inferences with widely-adopted devices such as wearables and smartphones. Our work makes contributions in the area of transfer learning and subject-specific representations, one of utmost importance in machine learning for health. Further, the proposed approach allows us to move beyond traditional intensity based metrics to truly capture inter-individual differences that would have otherwise been missed.

Remark: Our proposed methodology yields valuable user-level *embeddings* than can then be used in downstream classification tasks. Future work will explore how these individualized signatures may be used to better characterize large population cohorts, beyond what traditional, intensity based-metrics currently offer.

Finally, some interesting extensions to the transfer learning experiments would be to quantify the optimal number of hours or days we need for each user in order to accurately predict these health-related variables. Our current approach assumes that all temporal windows over the observation week for each user are aggregated resulting in a user-level embedding. This could have cost saving implications as well, given that large population studies like the UK Biobank [112] or the *All of Us* [391] procure wearable devices for large cohorts over long periods of time and therefore the shortest monitoring period would be beneficial. This shall be addressed in future work.

6.7 Conclusion

In this chapter, we proposed a novel *self-supervised* deep learning model to forecast HR responses from movement data using wrist-worn accelerometers. Through this architecture, our model learns user-level representations that can then be used for a variety of practical downstream tasks that are *personalized* to the users' unique physiology.

We evaluated this model in what, to the best of our knowledge, is the largest dataset of its kind, including over 1,700 participants with combined wearable ECG and wrist accelerometry for a week. Our model outperforms a set of benchmarks and we perform ablation tests to show the performance of different input modalities to the architecture.

In the forecasting task we found that including a single measure of RHR had significant impact on the error, indicating that an individual's continuous HR is greatly influenced by this marker. Moreover, we show that cyclical modeling of the hour and the month contributed to achieving the lowest error of $\sim$9 beats per minute in free living conditions. Nevertheless, even the model relying purely on acceleration (A/T) achieved competitive results ($\sim$12 beats per minute) outperforming other ML baselines.

We also introduced a joint *loss function* that acts as a regularizer to traditional MSE by using several quantiles of the predictive density of the model in order to capture the long-tails of HR observed in the real world. These joint losses outperformed single losses across all error metrics.

Lastly, we performed a set of health-related downstream, transfer learning tasks by aggregating the window-level features to user-level ones and showcasing the value captured by the learned *embeddings* through strong performance at inferring physiologically meaningful variables. For example, our models achieve an AUC of 70 for BMI prediction and an AUC of 80 for Physical Activity Energy Expenditure. By inspecting the embeddings we also noticed most outcomes improve in performance with higher latent dimensionality while some are invariant to its size, which needs further investigation.

CHAPTER 7

ADAPTABLE CARDIORESPIRATORY FITNESS PREDICTIONS FROM FREE-LIVING WEARABLE DEVICES

Publications

Part of this work is under review for publication in Heart.

Contributions

I planned this project and devised the analysis plan in collaboration with my colleague Dimitris Spathis and our supervisors. Dimitris and I have worked on the analysis collaboratively and I have written this chapter and the associated manuscript (work in progress at the time of thesis submission) which plan to submit for publication following its completion. The first part of this chapter also introduces association results in the same cohort from a study led by Tomas Gonzalez which I have also worked on as part of a team and is currently under review for publication.

7.1 Summary

Background: Low cardiorespiratory fitness (CRF) is a well-established predictor of cardiovascular and metabolic disease as well as all-cause mortality. An individual's CRF is usually determined through a VO_2max test. However, the setup of these tests is costly and burdensome making CRF assessments impractical in healthcare or large population studies. Due to the need for proxy measures, non-exercise, self-reported methods to estimate VO_2max have gained popularity in recent years, but their reliability is limited.

Methods: CRF (expressed as VO_2max) was estimated from a submaximal treadmill test. First, we examined the association of RHR (in beats per minute) and CRF from 11,059 participants that took part in the Fenland Study. Further, we exploited movement and cardiac signals extracted from wearable sensors in free-living conditions from these participants to derive a non-exercise model. Building on the findings from Chapter 6, we developed a deep neural network approach that leverages time-series sensor information extracted from multimodal wearable sensors to forecast individual's VO_2max in free-living conditions. Finally, we interrogated the ability of our deep learning models to adapt to change in CRF by evaluating it in a subsample of the Fenland population (*n = 2140*) that repeated the protocol after a median of 6 years (inter-quartile range: 5-8 years) from the original visit.

Results: The age- and sex-adjusted RHR (seated, supine, sleeping) association with CRF were: -0.26 (95%CI -0.27; -0.24), -0.31 (95%CI -0.33; -0.29), and -0.31 (95%CI -0.34; -0.29) ml O2·kg^{-1} · beat^{-1}, respectively. In the longitudinal analyses (Fenland I vs Fenland II), each 1-bpm increase in supine RHR was associated with 0.23 (95%CI 0.20; 0.25) ml O2·kg^{-1} · beat^{-1} decrease in CRF. Our wearable sensing non-exercise model yielded an R^2= 0.89 and RMSE = 1.61 (calibrated) and R^2= 0.67 and RMSE = 2.84 (non-calibrated). Further, preliminary results showed that our adaptive approach is able to forecast VO_2max in the Fenland II population with an R^2= 0.543, RMSE = 3.509 and correlation = 0.739.

Conclusion: Our findings show that RHR is a valuable biomarker of CRF and provides an adaptable non-exercise model for CRF predictions given the growing adoption of multimodal wearable sensors. The proposed model outperforms state-of-the-art non-exercise CRF models and regression based models using objectively measured physical activity, showcasing the value of habitual monitoring of physiological signals in free-living condition for these types of inferences. Additionally, we show that our deep learning models are adaptable to change and that the predictive performance improves when new sensor data is fed to the model. Our model could potentially be used in large population studies and health care settings, as well as to facilitate personal goal setting and remote patient monitoring of CRF.

7.2 Background

Cardiorespiratory fitness (CRF) is an important modifiable marker of cardiovascular health exhibiting a strong inverse relationship to the incidence of cardiovascular disease (CVD), type 2 diabetes, cancer, mortality and other adverse health outcomes across a number of published studies [481–484, 90, 485–487, 458]. Vast epidemiological and clinical evidence demonstrates that CRF is not only potentially a stronger predictor of mortality than well-established risk factors like hypertension, type 2 diabetes, high cholesterol or smoking, but that using CRF to complement these traditional risk factors significantly improves the precision of risk predictions for adverse CVD health outcomes [488, 483, 489, 490].

Beyond its implications in medicine, VO_2max is frequently used in athletics as an indicator of the endurance capacity of athletes, having strong predictive value for other sport-related variables [488]. The *gold-standard* measure of CRF is maximal oxygen uptake (VO_2max) which measures the maximal rate at which an individual can consume oxygen during exercise. Usually, VO_2max is measured through a maximal exertion incremental treadmill or ergometer test which require trained staff and expensive laboratory settings with specialized equipment [491, 492]. These tests require participants to reach volitional exhaustion while performing ventilatory gas analysis, which limits their scalability to large populations given the inherent risk they introduce. Furthermore, even for elite athletes, incorporating maximal tests can be inconvenient given strict training planning and the cumbersome nature of the tests.

Given these limitations, CRF assessments are not routinely performed in healthcare settings and indirect methods to estimate VO_2max through submaximal tests or non-exercise models have been developed. Submaximal models aim to predict VO_2max from continuous, incremental exercise at submaximal intensities [493]. These models are based on the parallel increase of heart rate (HR) and VO_2 consumption during exercise and assume a linear relationship between work rate and HR that holds at maximal intensities [494, 469]. Although these submaximal tests are valuable alternatives to maximal exertion tests, particularly for older and non-athlete populations, their scalability, cost, time consumption and potential risks associated to exertion still limit their use in clinical settings and applicability in large-scale population studies [495].

Resting heart rate (RHR) has not previously been employed as a proxy for CRF due to unresolved methodological challenges. RHR is known to be dependent on the physiological state at the time of measurement [496]. In clinical practice, the most common states are sitting upright during blood pressure measurement or lying supine during brief multi-lead ECG measurement. In free-living conditions, however, wearable sensors conveniently measure RHR in other states of rest, notably including sleep. It is unknown whether differences in RHR measured in these ways alter the relationship with CRF. It is also unclear to what extent the RHR-to-CRF relationship is affected by adiposity and physical activity, which have established associations with CRF [497, 498]. Quantifying the influence these modifiable factors have on the RHR-to-CRF relationship would strengthen researcher's ability to accurately use RHR as a

proxy measure of exercise capacity. Although some studies have described the longitudinal relationship between RHR and CRF [94, 499], there is uncertainty regarding how individual changes in RHR may reflect longer-term CRF changes in the general population.

Further, non-exercise models aim to provide an alternative for CRF assessment for widespread use in many healthcare settings and amongst individuals who cannot safely exercise. These models are usually regression-based and incorporate variables such as sex, age, body max index (BMI), the aforementioned RHR (although not at the scale of the association work presented on this chapter) and self-reported physical activity to infer VO_2max [500, 501]. However, model performance and reliability of is limited [502]. Wearable devices, such as activity trackers and smartwatches, provide a valuable opportunity for objective monitoring of physical behaviours, facilitating the capture of valuable physiological signals in a non-obtrusive, continuous manner. In particular, they enable continuous measurement of movement and cardiac responses, providing a scalable way to investigate the relationship between objectively measured physical activity, cardiac responses and CRF [503]. Although the use of these devices continues to grow, most of the derived variables in commercial wearable devices lack proper scientific validation and as such, their validity to make health-related inferences has been questioned [75, 504–506]. Specifically, VO_2max estimations using these devices have been shown to be particularly complex and at times unreliable [505, 507]. Whilst some commercial devices have shown stronger results than others, many rely on detailed activity intensity measurements, speed monitoring through GPS and require users to reach HRs close to their maximum capabilities monitored through chest straps, limiting the application to self-selecting, fitter populations [508, 509].

The relationship between submaximal HR during a variety of physical activities and VO_2max has been studied in laboratory settings [510–512]. However, these types of studies have rarely been conducted under free-living conditions and have mostly been limited to linear models imputing activity counts and daily steps [500, 513, 503, 514].

Chapter Significance: This Chapter is divided into two parts. In the first part, we aimed to assess the cross-sectional associations between different measures of RHR and CRF in a large population-based study of UK adults, as well as how body composition and physical activity may alter the association. Secondly, we examined whether longitudinal within-person change in RHR are associated with within-person change in CRF. In the second part, we aimed to substantially advance previous CRF non-exercise models by introducing a non-exercise model approach that leverages the information captured by the wearable sensors worn in free-living conditions. We used a deep neural network that utilises statistical features emerging from the wearable sensor data and demonstrated that these models yield better performance than traditional and state-of-the-art non-exercise models which do not use this free-living wearable data. Further, we showed that these models can be used to predict VO_2max using only participant information from the present.

7.3 Methods

7.3.1 Study description

The Fenland study is a population-based cohort study designed to investigate the independent and interacting effects of environmental, lifestyle and genetic influences on the development of obesity, type 2 diabetes and related metabolic disorders. Exclusion criteria included prevalent diabetes, pregnancy or lactation, inability to walk unaided, psychosis or terminal illness (life expectancy of ≤ 1 year at the time of recruitment) .

The Fenland study has two distinct phases. Phase I, during which baseline data was collected from 12,435 participants, took place between 2005 and 2015. Phase II was launched in 2014 and involved repeating the measurements collected during Phase I, alongside the collection of new measures. All participants who had consented to being re-contacted after their Phase I assessment were invited to participate in Phase II. At least 4 years must have elapsed between visits. As a result of this stipulation, recruitment to Phase II is ongoing.

After a baseline clinic visit, participants were asked to wear a combined HR and movement chest sensor Actiheart, CamNtech, Cambridgeshire, UK) for 6 complete days. For the first part of the study, data from 10,865 participants was included. These comprised all participants with complete relevant covariates and all three RHR measurements. For the second study, 11,059 participants were included after exclusion of participants with insufficient or corrupt data or missing covariates. A subset of 6,579 study participants (2,675 for the non-exercise model part of our work) returned for the second phase of the study, after a median of 6 years (inter-quartile range: 5-8 years). These participants underwent a similar set of tests and protocols, including wearing the combined HR and movement sensor for 6 days. All participants provided written informed consent and the study was approved by the University of Cambridge, NRES Committee - East of England Cambridge Central committee. All experiments and data collected were done in accordance with the Declaration of Helsinki.

7.3.2 Study procedure

Participants arrived at the testing facility after an overnight fast to complete baseline clinical assessment and questionnaires. Resting pulse rate was measured in a seated position while blood pressure was assessed three times at 1-minute intervals (Omron 705CP-II, OMRON Healthcare Europe, Hoofddorp, Netherlands). Seated RHR was computed as the mean of these three pulse rate values. At least one hour after arrival, RHR was measured with the participant at rest in a supine position using a combined HR and movement sensor (Actiheart, CamNtech, Papworth, UK) attached to the chest at the base of the sternum by two standard

ECG electrodes.[19,20] HR was recorded for 6 minutes and RHR was calculated as the mean HR measured during the last 3 minutes.

After their visit, participants wore the Actiheart *wearable ECG* which measured HR and movement recording at 60-second intervals [401]. The Actiheart device was attached to the chest at the base of the sternum by two standard ECG electrodes. Participants were told to wear the monitor continuously for 6 complete days and were advised that these were waterproof and could be worn during showering, sleeping or exercising. During a lab visit, all participants performed a treadmill test that was used to establish their individual response to a *submaximal test*, informing their VO_2max (maximum rate of oxygen consumption measured during incremental exercise) [430].

7.3.3 Cardiorespiratory fitness assessment

VO_2max was predicted in study participants using a previously validated sub-maximal treadmill test [515]. Participants exercised whilst the treadmill grade and speed were progressively increased across several stages of level walking, inclined walking, and level running. The test was terminated if one of the following criteria were met: 1) the participant wanted to stop, 2) the participant reached 90% of age-predicted maximal HR (208-0.7*age) [469], 3) the participant exercised at or above 80% of age-predicted maximal HR for 2 minutes.

7.3.4 Free-living wearable sensor data processing

Participants were excluded from this analysis if they had less than 72 hours of concurrent wear data (three full days of recording) or insufficient individual calibration data (treadmill test-based data). All participant HR data collected during free-living conditions underwent pre-processing for noise removal [402]. Non-wear detection procedures were applied to the Actiheart combined sensing data and non-wear periods were excluded from the analyses [516]. This algorithm detected extended periods of non-physiological HR concomitant with extended ($>$ 90 minutes) periods that also registered no movement through the device's accelerometer. This data was then modelled using a branched equation framework, resulting in physical activity energy expenditure (PAEE) (kJ/kg per day) time series [472]. We converted these intensities into standard metabolic equivalent units (METs), through the conversion 1 MET = 71 J/min/kg (3.5 ml $O_2 \cdot min^{-1} \cdot kg^{-1}$). These conversions where then used to determine intensity levels with ≤ 1.5METs classified as sedentary behaviour, activities between 3 and 6 METs classified as moderate to vigorous physical activity (MVPA) and activities > 6METs classified as vigorous physical activity (VPA).

Additionally, we derived a comprehensive set of features using the Python package *tsfresh* [517]. We performed ablation tests with and without these features during our experiments to evaluate their impact.

7.3.5 Metadata

Participants arrived at the testing facility after an overnight fast to complete baseline clinical assessment and to answer several questionnaires. For our analyses, we included ethnicity (binary: white, not-white), smoking status (never, former, current) and alcohol intake (units/week) from those questionnaires. Additionally, weight was measured with a calibrated electronic scale (TANITA model BC-418 MA; Tanita, Tokyo, Japan) and height was assessed with a wall-mounted, calibrated stadiometer (SECA 240; Seca, Birmingham, United Kingdom). Body composition was measured through dual-energy X-ray absorptiometry (DEXA) (GE Lunar Prodigy Advanced fan beam scanner, GE Healthcare, Bedford, United Kingdom) deriving fat, lean and bone mass measurements across all body regions [518].

7.3.6 Evaluation

To evaluate the performance of our deep learning regression models we computed root mean squared error (RMSE), mean sequared error (MSE), mean absolute error, Pearson correlation coefficient and the coefficient of determination (R^2). Here y and $\hat{y}$ are the measured and predicted VO_2max and $\bar{y}$ is the mean.

$$RMSE = \sqrt{\frac{1}{|N_{test}|} \sum_{y \in \mathscr{D}_{test}} \sum_{t=1}^{N} (y_t - \hat{y}_t)^2} \tag{7.1}$$

$$MSE = \frac{1}{|N_{test}|} \sum_{y \in \mathscr{D}_{test}} \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 \tag{7.2}$$

$$MAE = \frac{1}{|N_{test}|} \sum_{y \in \mathscr{D}_{test}} \sum_{t=1}^{N} |y_t - \hat{y}_t| \tag{7.3}$$

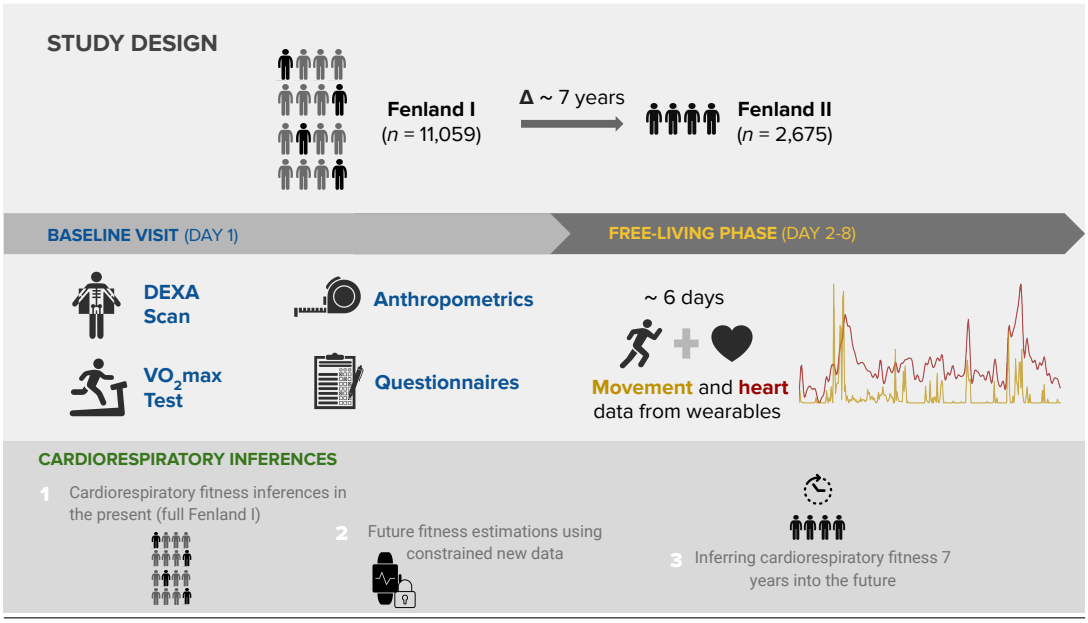
$$R^2 = 1 - \frac{\sum_{t=1}^{N} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{N} (y_t - \bar{y}_t)^2} \tag{7.4}$$

While MSE–RMSE put importance on the biggest errors, MAE gives the same importance to each error and is more robust to outliers. For the evaluation of the directionality of change, the evaluation metric is the Area under curve (AUC).

7.3.7 Statistical Analyses

For the first part of our work, the interrelationships of RHR measures were examined through linear regression. We used sex-stratified regression models to examine cross-sectional associations between RHR and CRF while adjusting for potential confounding factors. Model 1 adjusted for age; Model 2 additionally adjusted for demographic and lifestyle factors (ethnicity, smoking status, alcohol intake); Model 3 additionally adjusted for BMI. Subgroup analyses were performed as follows: Analyses were stratified by age group (less than 50y; 50-59y; 60y and above) and BMI (normal weight, BMI 18.5-25kg·m^{-2}; overweight, BMI 25-30kg·m^{-2}; obese, BMI above 30kg·m^{-2}). Descriptive statistics were reported as means and standard deviations, unless otherwise specified. For the non-exercise model part of our work we used the aforementioned evaluation metrics. For longitudinal analyses of the subsample with repeat measures of RHR and CRF, associations between within-person change in RHR and CRF were adjusted for baseline age, sex, RHR, CRF, and age at follow-up. Associations between RHR and VO2max were visualised with binscatter plots (5$\%$ bins) as shown in Figures 7.2 and 7.3. Statistical analyses were performed with STATA (Version 14.2; StataCorp, College Station, TX); a p-value of <0.05 was considered statistically significant.

Figure 7.1 Study and experimental design. (A) The Fenland study comprised of two assessment phases: Fenland I (n=12,435) and Fenland II including a subset of the Fenland I participants re-tested 6 years later. (B) During both phases participants underwent a variety of tests during the baseline clinic visit, including anthropometric measurements, questionnaires, DEXA scans and a submaximal VO_2max test. Following this baseline visit, participants were fitted with a combined activity and cardiac sensing device which they wore in free-living conditions for 6 days. (C) In this work we derive associations between RHR and VO_2max and introduce three sets of experiments for our non-exercise models: First, CRF in Fenland I is inferred leveraging free-living wearable data. Second, we demonstrate that even with scarce new information regarding the participant's future state, VO_2max can be inferred reliably. Finally, we show that our model is adaptable by re-training with the new wearable sensor information in the Fenland II assessment phase, yielding strong inference performance.



7.4 Results

Baseline measurements were collected from 12,435 healthy adults from the United Kingdom enrolled in the Fenland Study. A subset of these participants were assessed again after a median of 6 years (interquartile range: 5-8 years). Descriptive characteristics of the cohort's two phases used for analysis by sex and category are presented in Table 7.1. The mean and standard deviations (SDs) for each characteristic are presented in this table.

Table 7.1 **Characteristics of the study analytical sample: The Fenland I and II studies**

	Fenland I				**Fenland II**			
	Men (n= 5229)		Women (n= 5830)		Men (n=1448)		Women (n=1593)	
	N or mean	*% or sd*	*N or mean*	*% or sd*	*N or mean*	*% or sd*	*N or mean*	*% or sd*
Demographics								
Age (years)	47.70	7.57	47.66	7.36	54.51	7.10	55.08	6.81
Anthropometrics								
Height (m)	1.78	0.07	1.64	0.06	1.77	0.07	1.64	0.06
Body mass (kg)	85.85	13.83	70.54	13.92	85.88	14.23	70.22	14.54
BMI (kg/m^2)	27.16	3.97	26.17	4.97	27.25	4.20	26.17	5.26
FMI (kg/m^2)	44.32	12.21	45.67	10.84	N.A.			
Alcohol consumption								
<1/week	1328	25.4%	2262	38.8%	N.A.			
1-4/week	2823	54.0%	2746	35%	N.A.			
Almost daily	1009	19.3%	717	12.3%	N.A.			
Smoking status								
Never	2740	52.4%	3288	56.4%	N.A.			
Former	1757	33.6%	1883	32.3%	N.A.			
Current	685	13.1%	618	10.6%	N.A.			
Physical activity								
MVPA (min/day)	104.03	72.18	73.30	57.33	N.A.			
VPA (min/day)	12.52	18.15	6.28	11.04	N.A.			
RHR								
Supine RHR (bpm)	61.48	8.68	64.46	8.28	N.A.			
Cardiorespiratory fitness								
VO_2max per kg BM	41.95	4.61	37.44	4.73	44.04	10.33	36.40	8.02
VO_2max per kg FFM	58.08	7.52	58.53	8.78	N.A.			
Other								
Beta-blocker	22	0.42 %	15	0.26 %	67	4.62 %	58	3.64%

Data are in mean (SD) or n (%) unless otherwise indicated

7.4.1 RHR as a biomarker of fitness

Mean (SD) baseline RHR in the seated position, supine position, and during sleep were: 67.6 (9.8), 63.5 (8.9), and 56.9 (6.9) bpm, respectively, and correlations (Pearson r) between

modalities ranged between 0.65 to 0.81. On average, RHR was 3 bpm higher and VO_2max was 7.7 ml $O_2 \cdot min^{-1} \cdot kg_{-1}$ lower in women than men.

Table 7.2 **Association between resting heart rate and maximal oxygen consumption expressed per kg of total-body mass: The Fenland Study**. Reported values are beta coefficients (95% CI). Model 1: age-adjusted; Model 2: model 1 + ethnicity, smoking and alcohol adjusted; Model 3: model 2 + body max index (BMI) adjusted.

		Seated RHR		**Supine RHR**		**Sleeping RHR**	
		Men	Women	Men	Women	Men	Women
				Total sample			
Model 1		-0.27 (-0.29, -0.24)	-0.25 (-0.27, -0.22)	-0.33 (-0.36, -0.30)	-0.31 (-0.33, -0.28)	-0.28 (-0.32, -0.24)	-0.27 (-0.31, -0.24)
Model 2		-0.27 (-0.29, -0.24)	-0.25 (-0.27, -0.22)	-0.32 (-0.35, -0.29)	-0.30 (-0.33, -0.28)	-0.31 (-0.35, -0.28)	-0.31 (-0.35, -0.28)
Model 3		-0.23 (-0.26, -0.21)	-0.19 (-0.22, -0.17)	-0.29 (-0.32, -0.26)	-0.29 (-0.32, -0.26)	-0.26 (-0.30, -0.23)	-0.22 (-0.25, -0.19)
				Age stratified			
	<50 years	-0.26 (-0.30, -0.22)	-0.25 (-0.29, -0.21)	-0.34 (-0.38, -0.29)	-0.29 (-0.34, -0.25)	-0.35 (-0.41, -0.29)	-0.30 (-0.35, -0.24)
Model 2	50-60 years	-0.26 (-0.30, -0.23)	-0.25 (-0.28, -0.21)	-0.30 (-0.35, -0.26)	-0.31 (-0.35, -0.27)	-0.28 (-0.34, -0.22)	-0.32 (-0.37, -0.27)
	>60 years	-0.28 (-0.33, -0.22)	-0.24 (-0.30, -0.19)	-0.33 (-0.39, -0.27)	-0.32 (-0.38, -0.26)	-0.31 (-0.40, -0.23)	-0.33 (-0.41, -0.25)
	<50 years	-0.22 (-0.26, -0.18)	-0.20 (-0.23, -0.16)	-0.30 (-0.34, -0.26)	-0.25 (-0.29, -0.21)	-0.29 (-0.35, -0.23)	-0.21 (-0.26, -0.16)
Model 3	50-60 years	-0.23 (-0.27, -0.19)	-0.19 (-0.22, -0.15)	-0.27 (-0.31, -0.23)	-0.25 (-0.29, -0.22)	-0.23 (-0.29, -0.17)	-0.22 (-0.27, -0.17)
	>60 years	-0.25 (-0.31, -0.19)	-0.21 (-0.27, -0.16)	-0.30 (-0.36, -0.24)	-0.29 (-0.35, -0.23)	-0.28 (-0.36, -0.20)	-0.27 (-0.34, -0.19)
				BMI stratified			
	<25 kg/m^2	-0.30 (-0.35, -0.26)	-0.24 (-0.27, -0.20)	-0.35 (-0.40, -0.30)	-0.32 (-0.36, -0.28)	-0.34 (-0.41, -0.27)	-0.31 (-0.36, -0.26)
Model 3	25-30 kg/m^2	-0.21 (-0.24, -0.17)	-0.18 (-0.22, -0.14	-0.27 (-0.31, -0.24)	-0.22 (-0.26, -0.17)	-0.24 (-0.30, -0.19)	-0.17 (-0.22, -0.11)
	>30 kg/m^2	-0.18 (-0.23, -0.13)	-0.11 (-0.15, -0.06)	-0.22 (-0.28, -0.17)	-0.17 (-0.21, -0.12)	-0.19 (-0.26, -0.11)	-0.13 (-0.19, -0.07)

We explored several sex-stratified regression models to study the associations between RHR and VO_2max per kg body mass (Table 7.2, Figure 7.2). In models with only age adjustment (Model 1), RHR was significantly associated with VO_2max in women and men, irrespective of RHR modality. Associations were similar after adjustment for ethnicity, smoking, and alcohol use (Model 2). Further adjustment by BMI (Model 3) attenuated associations by about 10% for VO_2max per kg body mass. Associations between RHR and VO_2max were similar across all RHR modalities, especially in maximally adjusted models, but BMI and physical activity attenuated more of the relationship for sleeping HR in women (30% and 40%, respectively). We also analysed the association between RHR and VO_2max across age and BMI strata separately. Associations in these subgroups were similar in strength to pooled associations and remained statistically significant. Summary binscatter plots for these associations are presented in Figure 7.2.

In the subsample of participants with 6-yr repeat measures of supine RHR and VO_2max, mean levels of RHR and CRF were relatively similar to baseline values but with diverse individual change over time. The 5^{th} to the 95^{th} percentiles of change were -11.4 to 9.2 bpm and -10.1 to 10.8 ml $O_2 \cdot min^{-1} \cdot kg^{-1}$, respectively. Correlations between baseline and follow-up measures were high for both RHR (r=0.70) and VO2max (r=0.64). In longitudinal association analyses, each 1-bpm increase in supine RHR was associated with a 0.23 (95%CI 0.20; 0.25) ml $O_2 \cdot min^{-1} \cdot kg^{-1}$ decline in VO2max. Sex-stratified coefficients were -0.21 (95%CI -0.24; -0.17) and -0.25 (95%CI -0.28; -0.21) ml $O_2 \cdot min^{-1} \cdot kg^{-1}$ per 1-bpm increase in RHR in women and men, respectively. These findings are summarized in Figure 7.3.

Figure 7.2 Associations between RHR and VO_2max. Associations between resting heart rate and maximal oxygen consumption expressed per kg body mass, stratified by sex and adjusted for age. Top: Seated resting heart rate. Middle: Supine resting heart rate. Bottom: Sleeping resting heart rate. The Fenland Study (n=10,865). Each point represents 5% of data in the binscatter plots

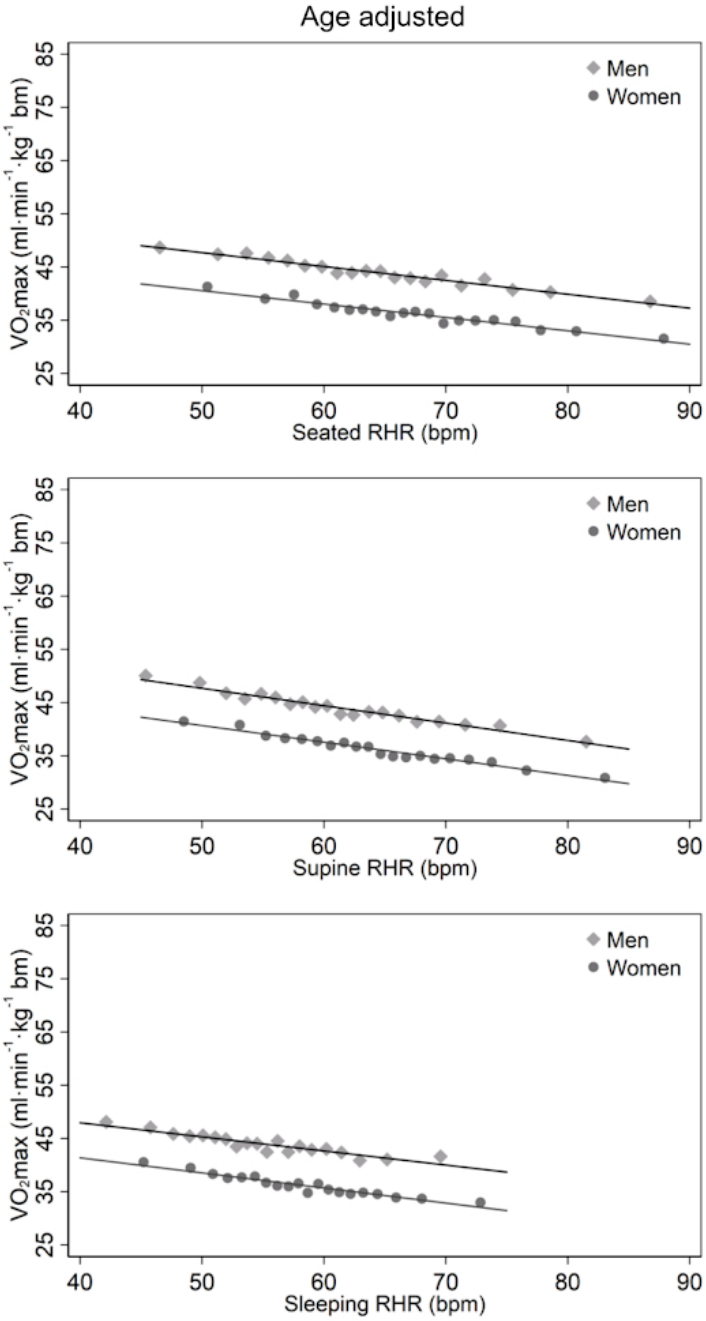


Figure 7.3 Associations between RHR and VO_2max over time. Association between 6-year change in supine resting heart rate and change in fitness, stratified by sex. Models were adjusted for follow-up time and baseline values of age, sex, RHR, and VO2max. Longitudinal subsample, the Fenland Study (n=6,589). Each point represents 5% of the data in the binscatter plot.

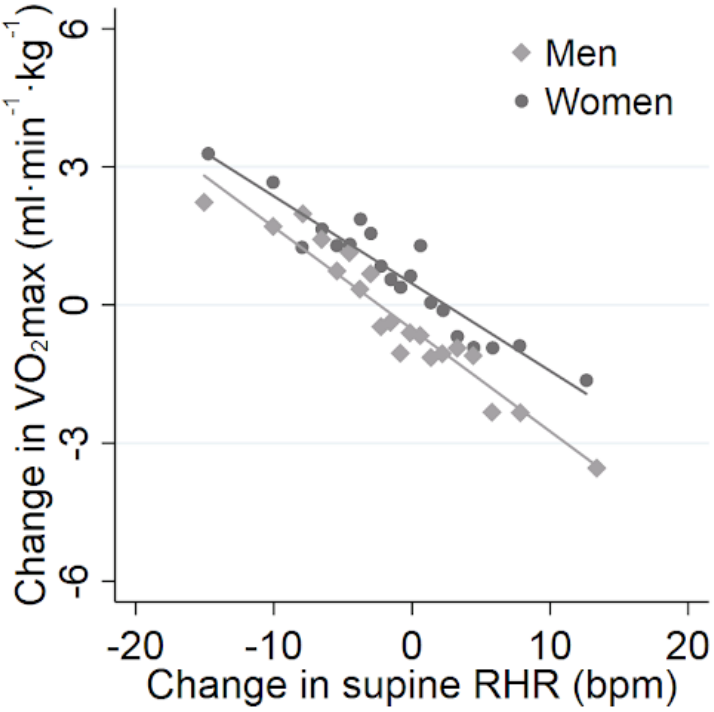


Table 7.3 **CRF inferences in the Fenland I cohort**. Inferences of VO_2max are presented in a sequential level of complexity. All results reported are for three layer convolutional neural network with regularization. Improvement upon linear regression was $\approx 2.5\%$ (R^2) across all inferences.

		R^2	Corr	RMSE	N(train/test)
Fenland I (predict VO_2max now)					11059(8384/2675)
	Anthropometrics Age/sex/weight/BMI/height	0.359	0.601	4.050	
	Anthropometrics + RHR Age/sex/weight/BMI/height/RHR	0.610	0.781	4.008	
	Sensors + Anthropometrics + RHR Acceleration/HR/HRV/age/sex/weight/BMI/height/RHR	0.649	0.806	2.999	
	Acceleration/HR/HRV/age/sex/weight/BMI/height/RHR\MVPA (non-cal.)	0.659	0.812	2.977	
	Acceleration/HR/HRV/age/sex/weight/BMI/height/RHR\MVPA (branch eq.)	0.931	0.964	1.341	

7.4.2 A deep-learning framework for cardiorespiratory fitness inferences from wearable sensor data

Based on the strong associations presented thus far in this chapter, we developed a more complex non-exercise model that utilises the wearable sensors used during the free-living protocol of the Fenland study.

To that end, we developed models (both regression and neural network based) that inputted a set of statistical features produced using *tsfresh* on the time-series sensor data as well as anthropometrics and RHR for each individual. We conducted ablation tests to study the influence of each subset of features. The results for the deep learning models are presented in Table 7.3 for the Fenland I population. These results were obtained using a simple neural network model consisting of 3 layers and regularization.

Our results show that using this free-living sensor data greatly added to the predictive capabilities of our model, resulting in improvements in terms of R^2, correlation and RMSE with each set of feature additions. Adding individual calibrated MVPA summary statistics yielded the best results (R^2 = 0.933). However, given that these branch equation models may not be available to use in any given cohort, we also report the non-calibrated results, which still yielded state-of-the-art results with an R^2 = 0.66. The calibrated result reflects the relationship between habitual PA and fitness with PA being optimally estimated with treadmill calibration, thus, if for completely protocol-free paradigms, it is more appropriate to look at the non-calibrated result. The performance improvement when compared to linear regression was of 0.2 points for all three metrics (R^2, correlation and RMSE).

7.4.3 Future cardiorespiratory predictions during the Fenland II assessment

Following our experiments in Fenland I, we evaluated our deep learning model on the subset of participants who returned for Fenland II. To do this, we trained on the subset of participants that returned for Fenland II (2140 training set, 535 test set) and performed transfer learning by using the first 3 layers of the network trained on Fenland I and fine-tuned with 3 extra layers in Fenland II. The results are presented in Table 7.4. Although still preliminary, the results are encouraging, yielding an R^2= 0.543, correlation= 0.739 and RMSE= 3.509.

Table 7.4 **CRF inferences in the Fenland II cohort**. Inferences of VO_2max are presented in a sequential level of complexity. All results reported are transfer learning results from pre-trained network on Fenland I.

		R^2	Corr	RMSE	N(train/test)
Fenland II (predict VO_2max future)					2675(2140/535)
	Anthropometrics + RHR Age/sex/weight/BMI/height/RHR	0.49	0.702	3.703	
	Sensors + Anthropometrics + RHR Acceleration/HR/HRV/age/sex/weight/BMI/height/RHR	0.529	0.731	3.561	
	Acceleration/HR/HRV/age/sex/weight/BMI/height/RHR\MVPA (non-cal.)	0.543	0.739	3.509	

Remark: This work demonstrates the promise held by wearable devices for the monitoring of CRF in large populations. First, our association work showed that RHR alone is a strong biomarker of CRF. RHR measures can be readily obtained from modern wearable devices (usually through PPG) while the participant is at rest or sleeping. Our non-exercise model shows that including time-series free-living data can greatly increase the forecasting abilities of non-exercise models. These models could allow for the ubiquitous monitoring of CRF without requiring laboratory visits or expensive testing equipment.

7.5 Discussion

The first part of our study showed strong interrelationships between measures of RHR (when seated, lying supine, and during sleep) and investigated their relationship with CRF (estimated maximal oxygen consumption expressed as VO_2max) in the Fenland study, a large population-based study of UK adults. Cross-sectional analyses demonstrated inverse associations between RHR and VO_2max that persisted across different RHR measurement modalities. Part of the association between RHR and VO_2max was explained by physical activity. In longitudinal analyses, within-person change in RHR was associated with within-person change in VO_2max. The magnitude of this association was similar to that observed cross-sectionally. RHR may therefore represent a suitable biomarker of CRF, and changes in factors determining CRF are paralleled by those that influence RHR.

Our association work is among the few existing studies to examine the influence of factors underpinning the RHR-to-CRF relationship, reporting significant inverse associations between RHR and CRF that are independent of age, sex or obesity. Together, these findings support the notion that RHR and habitual physical activity levels are intrinsically linked to exercise capacity. Thus, changes in CRF achieved through altered physical activity levels could be feasibly monitored with periodic RHR measurements.

Building on our association work, for the second part of this chapter we developed a non-exercise model that allowed for the inference of VO_2max by leveraging the information captured through the free-living sensor data recorded through the wearables that participants wore for $\approx$ 5 days. Our results showed that this information greatly improved the predictive capability of traditional non-exercise models and RHR alone, yielding an R^2= 0.89 and RMSE = 1.61 (calibrated) and R^2= 0.67 and RMSE = 2.84 (non-calibrated). Through ablation tests, we showed that the statistical features derived from the sensor data improved upon the performance of traditional non-exercise models using anthropometrics and RHR data only. These findings highlight the potential that capturing physiological signals from habitual PA has for fitness inferences. Further, we showed that through transfer learning, our model could be used to infer $VO_2max \approx$ 6 years into the future without the need for updated anthropometric information.

The work reported in this chapter has some notable limitations. We used heart response to a submaximal treadmill test to estimate VO2max, rather than a direct measurement. Even though our group has validated this approach [515], associations between RHR and CRF reported here may be influenced by residual error from the VO2max estimation process, which is largely dependent on reaching a percentage of age-predicted maximal HR. Additionally, VO2max estimated from HR response to submaximal exercise is unreliable in those taking medications such as beta-blockers. We excluded participants on beta-blockers, as well as participants who did not pass the medical screening for treadmill testing. Whilst this adds to the validity of our results, it means that the findings cannot be generalised to these individuals. This warrants future research to examine whether RHR could be used to clinically monitor CRF across the

lifespan. Similarly, the results presented for our ML experiments are still largely preliminary. Although the results are encouraging, deeper networks and other architectures should be tested as well as thorough hyperparameter tuning.

Future work should explore the use of time-series data directly on our model, for instance by replicating what commercial wearable devices seem to do when estimating VO_2max. These devices use 10-20 minutes of participant recorded activity (usually running) combined with GPS to calculate VO_2max based on how fast the user is running and the HR response associated to that. Further, we should explore the role of HRV in recovery more thoroughly as it has been shown to be an important biomarker of fitness as well. Finally, future work should also explore the role of uncertainty in these inferences, particularly given the fact that not all subjects will have the same amount of "valuable" data in these types of cohorts. For instance, these last proposed approaches might work well if the participants happen to exercise during the free-living period, but if the participant is mostly sedentary, the inferences might need to be reinforced by data from similar or matched participants in terms of anthropometrics, RHR and HRV measures to inform a better prediction.

CHAPTER 8

DIGITAL PHENOTYPING AND SENSITIVE HEALTH DATA: LESSONS FROM GENETICS & IMPLICATIONS FOR DATA GOVERNANCE

Publications

Part of this work is under review for publication in Science (Policy Forum).

Contributions

I led the writing and discussion pertaining to this chapter and the associated manuscript and worked collaboratively with a multidisciplinary group of scholars including computer scientists, public health and ethics researchers and lawyers.

8.1 Summary

Mobile and wearable technologies, as explored throughout this thesis, have the potential to transform healthcare by providing low-cost, objective measurements of digital phenotyping data at unprecedented scale. However, there are significant ethical and data governance considerations that must be addressed given the nature of the data collected and it's ubiquitous collection. In this final chapter we discuss the topic of data governance for digital phenotyping technologies, drawing on parallels and lessons from genomics research and highlighting areas which will require new governance frameworks.

Mobile and wearable devices, such as smartwatches and fitness trackers, increasingly enable the continuous collection of physiological and behavioural data that permit inferences about users' physical and mental health. Growing consumer adoption of these technologies has reduced the cost of generating clinically meaningful data. This can help reduce medical costs and aid large-scale research. However, the collection, processing, and storage of data comes with significant ethical, security, and data governance considerations. A complex ecosystem is developing, with the need for collaboration among researchers, healthcare providers, and a broad range of commercial entities across public and private sectors, some of which are not traditionally associated with healthcare. This has raised important questions in the literature regarding the role of the individual as a patient, customer, research participant, researcher, user and bystander when consenting to data processing in this ecosystem [519]. Here, we use the emerging concept of "digital phenotyping" [520, 521] to highlight key lessons for data governance which draw on parallels with the history of genomics research, while highlighting areas where digital phenotyping will also require novel governance frameworks.

Chapter Significance: This Chapter concludes the thesis by examining whether and how digital phenotype data ought to be regulated to ensure data privacy and security around user data, without unnecessarily halting important scientific progress in this area. This work draws from historical precedents in genetic research as well as providing with new frameworks and considerations unique to digital phenotyping.

8.1.1 Ubiquitous personal health data

Phenotypic traits are broadly defined as the observable characteristics of an individual that arise from the combined effects of their genotype and the environment. Analysis of phenotypes yields important insights across many fields of scientific research, from anthropology, where they have been used to improve our understanding of human life history, human evolution or human ecology, to health sciences, yielding advances in the identification of the genetic basis of disease and health-related traits, drug repurposing, and pharmacogenomics. Building on developments made through the collection and analysis of other phenotypic data, digital phenotyping can be defined as the "moment-by-moment quantification of the individual-level human phenotype using data from personal digital devices" [521, 17]. This process is often passive and allows for the quantification of the individual-level behavioural phenotype through personal digital devices such as mobile phones and wearable technologies [17]. Advances in these data collection tools have accelerated across both academia and industry, along with diverse applications in clinical and public health settings. These tools enable, at unprecedented scale, long-term phenotyping in free-living conditions with minimal subject attrition [521]. Further, while passive data collected and generated through mobile or wearable devices isn't without limitations, it overcomes some of the issues associated to traditional survey-based methods, such as self-report bias. Early examples include large-scale involuntary hand tremor

analysis via mouse cursor movement [522] and the use of Microsoft Bing search queries to detect neurodegenerative conditions [45].

While digital phenotyping offers many opportunities for researchers, clinicians and users such as the ability to make more informed decisions, ubiquitous and unobtrusive monitoring or to achieve better health outcomes, it also involves risks. Given the multidisciplinary nature of the field and the different levels of sensitivity of the data being collected, digital phenotyping interacts with a broad range of laws and governance regimes, ranging from medical and research ethics to contract law and data protection regulation. Moreover, there is an added layer of complexity associated to the alignment of regulatory frameworks that vary from country to country and are ill-adapted to the international nature and adoption of smartphones or wearables. As such, there is a risk that consumers will be insufficiently protected if they are exposed to digital phenotyping technologies which do not fall neatly within any existing consumer protection regime with an effective enforcement framework. Further, in consumer, unregulated scenarios, an important dichotomy arises between the user's motivation to use these technologies and the technology provider's incentives to collect, analyse, share or monetize the data produced by these users, turning into a prosumer commodity [523]. Historically, health-related data had been limited mostly to the academic and medical communities and regulated accordingly. However, the widespread adoption and continuous advancement of digital phenotyping tools and the types of inferences resulting from these technologies are paradigm shifting. A broad range of potential harms has been highlighted in the academic and policy literature, including unethical data collection [524], provision of inaccurate clinically relevant data [525], and discriminatory use of sensitive data, such as exclusionary insurance, employment discrimination, or unfair credit scoring [526, 527]. At the same time, a lack of regulatory clarity can also fail to provide commercial providers and researchers with the certainty required for ethical scientific research and innovation.

8.1.2 Digital phenotyping at scale

From a regulatory perspective, one difficulty posed by digital phenotyping is the use of data collected outside of a traditional healthcare context to make health-related inferences. There is already some scope for supervision by consumer protection agencies with relatively broad remit, such as data protection authorities (DPAs), in Europe under the General Data Protection Regulation (GDPR), and to an extent the Federal Trade Commission (FTC) in the US, for misrepresentations regarding health data governance. However, given the scale of the field, regulators continue to lack the cross-domain expertise and resources required to provide comprehensive oversight. Greater regulatory clarity on the categorisation and treatment of digital phenotyping data in different contexts, for example on the scope of EU GDPR definition of personal data concerning health, would allow agencies such as the FTC and DPAs to better allocate scarce resources. While governing digital phenotyping at scale may require new models of resource allocation and oversight, initial steps could focus on developing enforceable

industry standards, such as approved GDPR codes of conduct, to act as signals of consumer rights.

A 2019 study found that numerous mobile health apps still regularly failed to disclose processing of special category health data under the GDPR, instead providing only the more basic protections required for less sensitive data [11]. Even among prominent apps more prone to regulatory scrutiny, the complexity of terms and applicable regulations can prevent consumers from understanding the nature of their data being processed. For example, the Fitbit Privacy Policy treats the activity and fitness data that it directly collects as if it were non-health data, while noting that for any health data obtained from other sources, or other special category data under the GDPR, Fitbit will notify users and request separate explicit consent to process that data [1]. However, the same Privacy Policy separately notifies users of the possibility that a broad range of collected data, including exercise, activity, sleep, biometric, geolocation, and personally identifying information, may be collected. The result is that users may be unclear when accepting the Privacy Policy what forms of data Fitbit classifies as "health data" at that time and must trust that Fitbit will seek additional explicit consent to process this type of data.

Moreover, despite initially being developed as a consumer device, Fitbit and similar wearable devices are increasingly used to generate health-related insights in research settings. However, there are few agreed standards and minimal regulatory oversight on the necessary reliability of outputs for research and clinical purposes. Although classification as a "medical device" introduces requirements regarding validity of results, longitudinal reporting, notification to users of serious health concerns, and improved reporting, only some device manufacturers have elected to seek classification as a medical device (i.e.: Apple Watch's electrocardiogram (ECG) App obtained De Novo FDA clearance in the US, and is classified as a Class II medical device), and often only for some device functions [528]. Moreover, these designations tend to change across the globe and as a result, consumers may be under the impression that the device has been subject to a greater degree of regulatory scrutiny than is necessarily the case. When monitoring data important to their health, consumers in these circumstances may place undue weight on potentially unreliable outputs, with the potential for misinformed self-diagnosis or behavioural change.

Companies utilising digital phenotyping regularly make unilateral changes to their terms of service, and privacy disclosures are frequently inadequate, underscoring a lack of protection for personal data and user privacy [529]. Undisclosed sharing of digital phenotyping data, including linkable identifiers, is prevalent [530]. Inconsistent regulatory oversight, unclear terms and conditions, and failure to disclose data sharing and secondary use can limit the ability for health care professionals to recommend otherwise beneficial apps in fields such as mental health care [530]. In the research context, the GDPR adopts lower protections for data subjects where the purpose of processing is solely for statistical or scientific research purposes. Despite

[1] Fitbit Legal Privacy Policy. Fitbit.com. (2020). Retrieved 20 March 2020, from https://www.fitbit.com/us/legal/privacy-policy.

submissions from some concerned groups, such as the BioMolecular Resources Research Infrastructure-European Research Infrastructure Consortium (BBMRI-ERIC) regarding the need to define "scientific research" to exclude some forms of commercial processing [531], the GDPR was passed to also allow commercial providers to use this research exemption to process sensitive personal data without consent (though still subject to EU Member State law, technical safeguards, and research ethical standards). Clear policy guidance on data sharing practices in these instances is critical to maintaining public trust in scientific digital phenotyping research and enabling the use of these methods for clinical care.

8.1.3 The risks of digital phenotyping: lessons from genetics

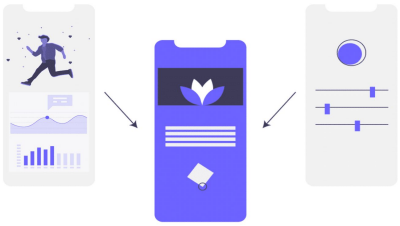
When considering improvements to the framework for digital phenotyping governance, there are valuable precedents from an earlier wave of health technology innovation. Advances in genotyping techniques, particularly from the 1990s onwards, created an extraordinary opportunity to better understand human health. At the same time, the sharing of the sensitive individual-level health data required for scientific advances created the need to develop new standards, policies, and regulations for genetics and bioinformatics research. This allowed policymakers to enact measures such as obligatory genetic counselling, requirements for validity of results, informed consent, and chain of custody procedures, which flowed from iterative resources such as the Bermuda Principles, Oviedo Convention, genetic testing protocols, and Council for International Organizations of Medical Sciences (CIOMS) guidelines.

To advance human health and infectious disease research, cross-border data sharing has become essential in genomics, leading to the creation of a variety of genomic data resources. These databases are mainly constructed by and for publicly funded scientific and medical researchers and their institutions. They range from being completely open, like the BRCA exchange, ClinVar, and Genome Aggregation database [532–534], to having regulated-access like the European Genome-Phenome Archive, the database of Genotypes and Phenotypes (dbGaP), and the Human Gene Mutation database [535–537]. Potentially instructive models to draw on for digital phenotyping data include controlled/managed-access models, data access committees, data safe havens, dynamic and tiered consent, differential access, explicit open-access consent, and portable legal consent. In particular, dynamic and tiered consent models are readily applicable to areas of digital phenotyping, where the sensors for data collection tend to be associated with a consumer device, such as a mobile phone, which could more easily support a user-friendly interface for dynamic consent models. Through collaboration across researchers, commercial providers, and regulators it may be possible to leverage these technology platforms to further improve the delivery and application of data management solutions developed in genomics.

A series of studies has demonstrated the challenges for researchers of fully anonymising data (including in controlled-access databases such as the dbGaP), observing data subjects' bounded

consent on collected data, and delivering clinically valid and meaningful data in a direct-to-consumer (DTC) setting [538]. While the lessons learned from genomics in these areas can assist with approaching digital phenotyping data governance, we must also be mindful of the important differences between genotypic and phenotypic data. Whilst genotypic data solely comprises genetic code, digital phenotypic data is extremely diverse. As a result, the data collected under the umbrella of 'digital phenotyping' may give rise to a broader range of possible harms.

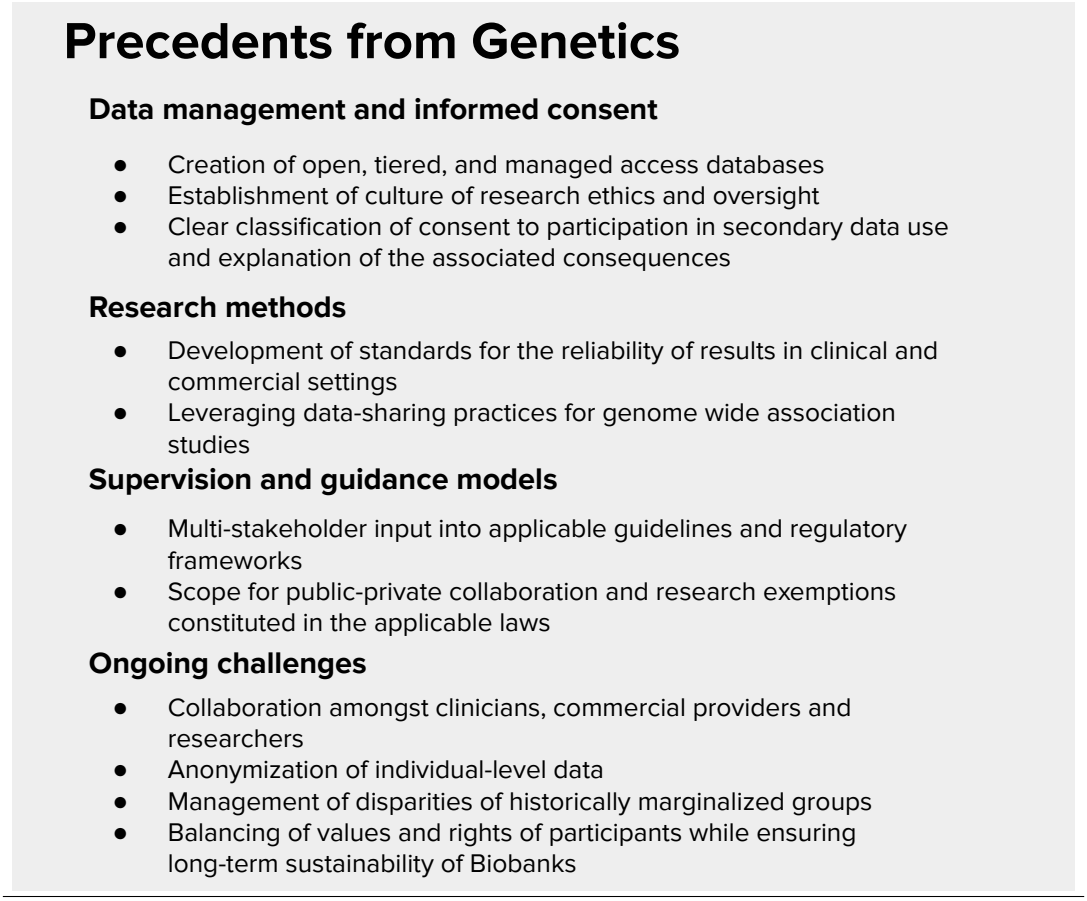
Figure 8.1 Differences between genotyping data and digital phenotyping data. Digital phenotyping can never be said to be "complete", because new data is generated continuously to reflect changing patterns of user behaviour. Although sophisticated data analysis often requires considerable infrastructure and expertise, the cost of processing and analysing each additional data point is usually negligible.



GENOTYPING DATA	DIGITAL PHENOTYPING DATA
Single dataset	Complex, multi-layered/modal datasets
Collected once	Collection ongoing
Unchangeable/immutable	Changeable/mutable
Often knowledge of collection	Sometimes knowledge of collection
Non-negligible cost/expertise	Negligible cost/expertise
Marginal benefits from triangulation	Substantial benefits from triangulation

Though there are a number of differences between genotyping and digital phenotyping data, the disparity is especially apparent in the manner of collection. The negligible costs of most types of digital phenotyping data collection means that it is efficient to aggregate large digital phenotyping datasets from different users who may be unaware of their inclusion in a dataset where personal devices upload data by default to a centralised server. Even in cases where explicit agreement is sought in terms of service for data collection, the terms may not reflect the nature of consent for research or secondary use, and complex terms and conditions can result in user fatigue and a "tick-box" approach, meaning that users are less likely to provide fully informed consent [529]. Moreover, the vast majority of digital phenotyping data arises from commercial products, where the role of this data and the associated research is at least in part to support a business model. Most of this data is therefore not used to produce pure public goods or knowledge and is not freely available under existing governance frameworks for proprietary data.

Figure 8.2 Building on developments in genetics to establish a path for digital phenotyping.



The current fragmented approach to regulatory oversight, classification of data for the purpose of identifying the applicable laws, and varying data governance practices lowers user trust in digital phenotyping and limits potential medical research. While the precedents of considered regulation and multi-stakeholder collaboration in genomics should inform developments in this field, it is also important to improve on these models where possible, and address aspects of digital phenotyping which require novel solutions.

While genetics databases have generally tended towards releasing aggregate data, with an inverse relationship between accessibility and individual-level data, the unique collection and delivery platforms of digital phenotyping may offer new models for data management and informed research participation. For instance, these technologies include the means to reduce the current practice of centralised data consolidation for the purposes of extracting value. Through privacy-preserving, decentralized methods like federated learning and zero-knowledge proofs [539, 540], users could maintain sole custody of their data [541]. These methods also enable model sharing, as opposed to data sharing, which could allow for more seamless cooperation between corporations and academic or public sector institutions. Advances in differential privacy techniques could also address this issue by collecting and aggregating information about groups of users' habits and behaviours while not sharing data from individual

users, potentially compromising their privacy. Similarly, from a consumer-facing perspective, digital phenotyping technologies could enable innovation in dynamic consent and through modern user interfaces and devices.

Given the nature of the data collected, laying the foundations for responsible data governance and providing reliable, well-validated and contextualized outputs will be critical to building trust and enabling the development of the digital phenotyping field. Mobile and wearable technologies have the potential to transform healthcare by providing low-cost, objective measurements of physical, cognitive, emotional, and social behaviours at unprecedented scale. By building on advances in genetics to promote good governance and user trust, digital phenotyping techniques could be used to understand disorders, aid diagnosis, improve patient monitoring and provide personalised coaching and interventions in the near future.

CHAPTER 9

DISCUSSION AND IMPLICATIONS

9.1 Summary of the aims and rationale of this thesis

The objective of this thesis was to advance understanding of multimodal sensing of physical behaviours in free-living conditions by deriving new insights from this type of data. These insights, in turn, provide a basis for better understanding human behaviour and its consequences for health and wellbeing.

The studies that comprise this thesis addressed this central aim through 4 sub-aims, outlined in detail in Chapter 1, relating to important gaps in the current literature. These comprise:

- Aim 1: To estimate sleep and sleep stages in free-living conditions through multi-modal wearable sensing
- Aim 2: To explore the use of wearable sensors to better understand and characterise physical activity
- Aim 3: To leverage multi-modal free-living sensing information to infer meaningful physiological characteristics
- Aim 4: To summarise and provide recommendations for the future management of mobile sensing data and digital phenotyping

The proliferation and growing adoption of wearable devices in both research and commercial settings has enabled an outstanding opportunity to measure physical activity and sleep at scale. While many heuristic models were developed during the 90s to classify sleep during the night using actigraphy devices, these algorithms rely on brand-dependent activity counts and were derived to be used during a given sleep window only, limiting their generalisability. These algorithms were derived to be used only during the night-time period, severely limiting their applicability to studies without sleep diaries or expert annotations and leaving an important gap on methods to identify sleep in free-living conditions to be addressed. Similarly, although commercial devices such as those produced by Garmin or Fitbit give sleep estimates, their work has yet to be validated and their models are also proprietary. Aim 1, which looked at estimating sleep and sleep stages in free-living conditions through multi-modal wearable sensing, was addressed through the work in Chapters 2-4. Together, these studies aimed to provide more accurate and scalable possibilities for sleep inference using multi-modal devices.

First, we explored how to leverage new multimodal sensing technologies to track sleep outside of laboratory environments. We began by providing an in-depth overview of state-of-the-art technologies and analysing their strengths and limitations (Chapter 2). We then provided a holistic analysis of sleep and sleep stage classification using conventional heuristic models as well as machine learning and deep learning techniques in the largest dataset available to-date with multimodal sensing and ground truth for sleep and sleep stages. These experiments highlighted, for the first time, the strengths and limitations of the generation of wearable

sensors when predicting sleep stages at different levels of granularity and which sensor features where most important in those inferences. We also showed how an ensemble model can be used to improve upon conventional deep learning models (Chapter 3). Similar models could be used to drive even stronger results in multi-class settings and may benefit from the inclusion of domain knowledge regarding the transition probabilities across the different sleep stages. We then developed a device and sampling rate agnostic algorithm to classify sleep periods based on HR alone, showcasing its results in four datasets. These included two free-living datasets, where the performance of the algorithm was evaluated across multiple nights on the same participants showcasing it's ability to capture intra and inter-individual variation, and in two datasets where the evaluation was done against PSG, including a study where the evaluation was carried using a commercial device (Apple Watch) (Chapter 4). The strength of this method lies in the fact that it does not require expert annotations or a sleep diary and can effectively track unconventional sleep schedules, such as shift work and naps, something that no prior algorithm had been able to do.

Under Aim 2, in Chapter 5, we explored how pitch and roll can be used to describe physical behaviours beyond what traditional intensity metrics from accelerometers offer. We found that sedentary and physical behaviours are better characterised through these derivations than when using intensity based-measures alone. For instance, our group has recently used pitch and roll to showcase that these measures can be used to differentiate between lying, sitting, standing, and moving in a study of hospitalized older patients [114]. Similarly, as explored in Chapter 4, in the absence of HR data, pitch and roll signals from accelerometers can provide valuable indications of when a participant is not changing posture frequently, which can help infer sleep.

Under Aim 3, we used self-supervised learning to map movement to HR data, leveraging the vast amounts of unlabelled time-series data that characterise large epidemiological studies including objective monitoring through wearable sensors (Chapter 6). Through this architecture, we obtained valuable embeddings which could then be used to infer physiologically meaningful information at a participant level (BMI, age, sex, VO_2max, etc). These findings were important because they provide valuable evidence that the pre-training task generated embeddings that learned personalized features for each participant fitted with these combined sensing devices. These results showcase a transition from traditional summary metrics that capture when and how much of a behaviour has taken place, such as exercise, and are a stepping stone towards capturing something more meaningful, such as the consequences of that exercise. As such, this method represents one of the first approaches that enables capturing individual characteristics by leveraging contiguous multimodal wearable signals. Building on these findings, we devised a deep learning approach that leveraged multimodal sensor data to infer an individual's CRF (VO_2max) in the present, as well as to forecast individual's fitness 6 years into the future (Chapter 7). Our findings represent the first of their kind on the inference of CRF, given the size of the cohort and the future prediction aspect of our work. In sum, these studies provide an important step towards the development of scalable computational techniques to measure,

characterise, understand, infer and enhance human behaviour and health through multimodal wearable sensors.

Finally, under Aim 4, we outlined the implications for data governance that arise given the advancements in digital phenotyping technologies such as those presented throughout this thesis (Chapter 8). While harnessing the widespread adoption of mobile devices to generate clinically meaningful data can help reduce medical costs and aid large-scale research, the collection, processing, and storage of data require careful consideration of privacy, security, and data governance. In this final chapter, we highlight the need for a new governance framework with regards to digital phenotyping data, which draws on the history of genomics regulation to provide clarity and protection for consumers, researchers, and commercial providers.

9.2 Discussion and main contributions

Traditionally, healthcare has primarily taken the form of therapy and drug treatments given once an individual presents with signs of ill-health. However, digital health, made possible by wearable devices and a new era of the "quantified self", is an emerging option that allows the continuous, longitudinal monitoring of behaviour and physiological markers at scale. Not only is this likely to be of research benefit in understanding the health causes and consequences of physical behaviours but it could also ultimately facilitate health monitoring and the early identification of disease, leading to better planning, earlier intervention and better outcomes. For example, early behavioural markers of Alzheimer's Disease can be identified through changes picked up by wearable sensors before clinical symptoms become apparent [542]. Despite their potential utility, current wearable technologies have sizeable challenges to overcome such that they can be reliably and usefully implemented in research and health monitoring. These primarily relate to leveraging the multimodal nature of modern devices and to the validation of inferences, generalizability and the development of an understanding of how measurement errors may affect research conclusions or impact decision-making in clinical settings.

Some of the most challenging problems surrounding wearable device technology in healthcare and research settings relate to the reproducibilty, transparency and adaptability of algorithms across datasets. Most commercial devices use proprietary algorithms which make reproducibility and validation challenging. In academic settings, algorithms are typically derived in small, constrained populations often meaning that their performance is not sustained when deployed in a more diverse set of users or settings. Another form of variability in wearable and mobile sensor data is caused by the positioning of the sensors on different anatomical locations [543]. As such, some models fail to be robust when exposed to the real-world heterogeneity of wearable and mobile sensor data. Further, while multimodal approaches have shown great promise in laboratory-based HAR tasks [116, 46] the use of multimodal wearable devices for health-related inferences is still at a nascent state. We believe that the impact of these combined sensing approaches could amount to a revolution in the objective monitoring of physical behaviours of similar magnitude to the one experienced through the introduction of pedometers and actigraphy devices which enabled, for the first time, a move beyond self-reported measures. For instance, they will allow a move beyond traditional intensity-based summary measures and provide individual signatures which encompass much richer individual physiological information and are truly personalized.

Throughout this work, we introduce a set of studies that shed light on how these multimodal sensor approaches can augment single-modality inferences by capturing more complex phenotypes and validate these inferences in large and diverse cohort studies.

9.2.1 Main contributions

In this thesis, we have introduced a number of studies that demonstrate how multimodal wearable sensors can be used to infer physical activity, sedentary behaviours and sleep in free-living conditions by leveraging signal processing and machine learning methods. Further, we have shown how deep learning and, in particular, self-supervision may enable vast amounts of unlabelled data to be leveraged. This could then be used to make personalised inferences of an individual's health and fitness levels.

Chapters 2, 3 and 4 provided an overview of how new sensing technologies can be used to effectively monitor sleep outside of laboratories, in free-living conditions. Through this work, we showed that although wrist-worn wearable sensors can be used reliably to track total sleep time and sleep onset and offset reliably, their performance when tracking sleep stages, even when using multimodal cardiac and movement sensing is more limited. Further, we showcased that HR sensing alone can be used to infer sleep windows in the absence of sleep diaries. These findings are of particular importance given that HR sensors are now common in wearable devices, many large cohort studies do not have sleep diaries and, even when these are available, they are prone to recall bias [85]. Further, our method can be used in non-habitual sleepers, such as shift workers, and to track naps, both valuable additions to the objective monitoring of sleep in free-living conditions.

For our work on multimodal sleep sensing we found that whilst the performance of multistage classification into Wake, NREM and REM was acceptable, as more granular sleep staging was introduced, the models struggled to differentiate the most common NREM substage (N2) and identify REM sleep for some participants. Hence, in future, commercial companies presenting these types of more detailed inferences must caveat them and produce validation studies that showcase their performance to avoid misleading customers. An alternative potential solution might lie in focusing on the transition to bed, something which might be more actionable and result in people getting a better night's sleep and more regular sleep schedules, which also have important implications for health and well-being. In parallel to the work described in this thesis, we have developed a number of open-source multimodal wearable sensing processing tools in Python that have been made available through Github.

Chapter 5 presented a descriptive epidemiological study showcasing the value added by including postural features in physical activity studies that use triaxial accelerometers for better characterisation of human behaviours. Inclusion of pitch and roll allowed for a more comprehensive understanding of human activity, beyond traditional intensity metrics derived from these accelerometers. These metrics can be easily obtained from modern triaxial accelerometers and have proven beneficial for the characterization of human behaviours. Leveraging these insights, pitch and roll were then used to derive a modified version of the HDCZA method to characterize limb immobility and inform sleep periods in Chapter 4. Similarly, pitch and roll have recently also been used by our group to differentiate between lying, sitting, standing,

and moving in a study comprising hospitalized older patients [114]. Together, the studies contribute to the literature by advocating for the inclusion of pitch and roll in inferring physical behaviours.

Chapter 6 introduced a self-supervised architecture that "mapped" activity obtained from a triaxial accelerometer to HR. Through this auxiliary task we hypothesised that the model would learn to capture useful, personalised representations that reflect individual characteristics. Thus, we tested the information carried in those representations through a variety of downstream transfer learning classification tasks that related to individual characteristics. Through these tasks, we showed that the embeddings emerging from the self-supervised tasks could be used to infer age, sex, BMI, height and VO_2max (among other characteristics). To the best of our knowledge, this is the first time a self-supervised architecture has been used used to generate physiological signatures through these embeddings. information captured by modern multimodal sensors could hold even greater promise. Current single modality sensors can tell us when and how much one particular behaviour, like exercise takes place. Multimodal sensors coupled with deep learning as we showed in this chapter could inform the effects that these behaviours have in the user's physiology, helping inform better decisions about our bodies and lifestyles. This work inspired us to build a predictive model of CRF using deep learning in Chapter 7. In this final analysis chapter, we found that resting and sleeping HR have a strong inverse association with CRF and devised a model that leveraged sensor information as well as some basic anthropometric information to infer current CRF (VO_2max). Further, we showed that our model could be used to infer VO_2max 6 years into the future using scarce new information (showing its performance with updated age and age, weight and BMI) and that the performance improved further when retraining with new sensor information collected 6 years in the future. This work has important implications for the role of wearable devices to monitor CRF at scale. Indeed, the ability to ubiquitously and reliably infer CRF using wearables could have important implications for preventative medicine, given its strong associations with cardiovascular and metabolic health.

Finally, this thesis concluded with a chapter that explored data governance considerations for the type of data generated throughout this thesis, termed digital phenotyping (Chapter 8). The chapter used precedents from the field of genetics to highlight some valuable lessons from the past and areas that are unique to this type of data. Mobile and wearable technologies have the potential to transform healthcare by providing low-cost, objective measurements of digital phenotyping data at unprecedented scale. Given the nature of the data collected, it is important that the sequencing of the digital phenome follows data governance principles that guarantee the rights of users, prevent misuse of data and promote trust in the rapidly evolving digital health ecosystem.

Together, the studies that comprise this thesis demonstrate that multimodal digital phenotyping technologies, in particular wearable sensors, can be used to accurately infer sleep, fitness and other physical behaviours in large-scale populations, validating these results against gold-

standard measures in diverse cohorts. Further, the results presented in this thesis point towards several interesting future opportunities for research, some of which we shall outline here, concluding the thesis.

9.3 Future directions

9.3.1 Domain adaptation for mobile and wearable sensing

One of the most interesting and important challenges that remains to be addressed in the field of wearable sensing is transfer learning and domain adaptation. This concept was featured on the introduction of this thesis and entails that a model trained on a particular device would be seamlessly applied to other devices that may have a different sampling rate or be slightly different in nature. This is particularly important when considering that currently, models that are learned using particular devices or in particular populations may then be deployed in completely different data distributions, halting the value of those inferences. For instance, a machine learning model derived in a particular population (distribution) may not translate well to a different population (different distribution) which could yield misleading results. This is one of the reasons behind why we decided to build a device and population agnostic HR approach (Chapter 4) as opposed to using one of the trained models for Chapter 3 and apply it to any given population without appropriate domain adaptation.

Ideally, a model trained on a particular device would be seamlessly applied to other devices that may have a different sampling rate or be slightly different in nature. To do so, we turn to *domain adaptation*, a discipline that seeks to develop learning algorithms that can be easily ported from one domain to another, from one device or brand to another [100]. Domain adaptation techniques have shown strong results in natural language processing, image and video classification tasks [101, 100, 102]. More recently, these techniques have been applied to wearable and mobile devices in the context of human activity recognition tasks [103–105]. Most existing approaches implement this philosophy of alignment by minimizing a measurement of distributional discrepancy in the feature space, often some form of maximum mean discrepancy (MMD) [106, 107], or a learned discriminator of the source and target as an approximation to the total variation distance [108].

One potential approach for domain adaptation in wearable and mobile sensing data would be to use *self-supervised* pre-training with the objective of aligning the source and target domains. Most wearable device datasets, particularly those collected in free-living conditions, lack proper labels as they are expensive and burdensome to achieve. Indeed, if the target domain dataset had labels of the same nature as the source domain, a normal classification task would suffice. However, due to the lack of labels in the vast majority of these datasets one could introduce a set of *self-supervisory auxiliary tasks* which create their own labels for the data.

By using multiple self-supervised tasks pertaining to different aspects of the wearable sensor's signals and training both domains together with the original task on the source domain, we would anticipate producing robust, valuable and well-aligned representations that can then be used for downstream classification or regression tasks on the target domain.

9.3.2 Improved human activity recognition through semi-supervised self-training of wearable device data

Mobile and wearable computing's key component in general, and for context awareness in particular, is the automated assessment of what a user may be doing at any given time, which is commonly referred to as Human Activity Recognition (HAR) [544]. The biggest challenge in using artificial intelligence techniques for HAR scenarios is the absence of large scale, *labelled* training datasets. Similarly, ground truth labelling is often of mixed quality or ambiguous, which complicates training as well as validation of models built for HAR purposes. New approaches like camera-based annotations and free-living GPS self-annotations have been implemented to overcome the non-naturalistic environments of most conventional HAR experiments. Sensors used for HAR purposes record temporal data, thus, HAR models need to solve a *dual problem:* localise the contiguous portion of temporal data that is relevant to the activity recognition problem (*segmentation*) and to *classify* those extracted segments into particular classes. This represents an important challenge when compared to other fields where deep learning techniques have been very successful, such as *computer vision* or *natural language processing*. In these fields, massive amounts of labelled data (i.e.: Imagenet, CIFAR, or the English Wikipedia) enable the generation and testing of new deep learning models. Currently, there are very few datasets that would enable this type of work in HAR, and those that are available are very limited given the important privacy issues that they raise and annotation costs associated with these type of tasks at scale. Hence, most approaches to date have been limited by the size and non-free living nature of these labelled datasets and there is a lack of reference models that can be applied to natural, free-living conditions [450, 545, 546].

Semi-supervised learning approaches are a potential solution to the current limitations in HAR as they could leverage unlabelled wearable data in free-living conditions present in large population studies like the UK Biobank and combine them with the small amounts of labelled data from smaller HAR studies [112]. Through unsupervised representation learning, semi-supervised models take full advantage of the large diversity of activities and behaviours present in large unlabelled data. Then, those learned representations can be fined tuned with a smaller, labelled dataset through a supervised task that can then be used to label the full dataset. In UK Biobank, the representations learned through unsupervised learning would not only be rich with regards to the size of the population (100,000 participants), but also in their nature (free-living, natural behaviours versus traditional constrained environments).

Self-supervised learning is a training paradigm that leverages the intrinsic structure present in the input signals [65]. These models make use of large scale unlabelled data by learning objectives in order to *get supervision from the data itself*, using supervised loss functions. Through this process, objectives are computed from the signals themselves by applying known transformations to the input data. Most importantly, the intermediate representations capture semantic and structural meaning that can then be exploited for a variety of downstream tasks. While there might be conceptual overlaps with unsupervised learning, its goal is primarily to learn a good geometrical structure of the data using the reconstruction loss (how well we can compress and decompress the input) so that the result can be used mainly for clustering purposes. On the other hand, self-supervised learning's goal is to leverage relationships between the input data using supervised losses (classification or regression). This new paradigm has been successfully applied in filling the blanks for image datasets [66], next word prediction [67], video frame order validation [68], and more recently to small-scale activity data [69]. In BERT [67], for example, the task is to predict the next word given the past sequence. BERT outperforms any other language modelling method by adding two auxiliary tasks within it's architecture, both of which are based on self-generated labels. On the other hand, the pre-trained task of video frame order validation [68] improved the performance on the downstream task of action recognition when used as first step. Similarly, in Chapter 6 of this thesis, we used a pre-training framework that employed self-supervision to generate embeddings that were then used in downstream classification tasks.

A possible solution lies on the use of *self-supervised* approaches that effectively utilise the power of unlabelled data combined with the use of a *semi-supervised teacher/student* paradigm which would use smaller, labelled datasets to fine-tune the model with true labels. We recently submitted a piece of work that laid the foundations for future work in UK Biobank and other large unlabelled datasets. The proposed approach, coined *SelfHAR*, is a self-supervised model that effectively learns to leverage unlabelled mobile sensing datasets. Our approach combines teacher-student self-training, which distil the knowledge of unlabelled datasets whilst allowing for implicit data augmentation, and multi-task self-supervision, which learns robust signal-level representations by predicting distorted versions of the input. Early results showed that this architecture yields state-of-the-art performance over supervised and previous self-supervised approaches, with up to 12% increase in F1 score using the same number of model parameters at inference.

9.3.3 Robust methods and data science tools for large-scale observational studies

Throughout this thesis we have used multiple cohorts that included large-scale wearable data to gain actionable insights on physical behaviours, health and fitness traits. In future, the methods used to derive these measures could be leveraged in large population studies for analyses exploring their associations to health and disease. Ultimately, this may be used to

inform health guidelines and recommendations. However, given the observational nature of most of these large cohort studies it is important to be able to rule out potential confounding factors and other potential alternative explanations for any associations reported, including inaccurate measures of the variables included in these studies to derive the variables of interest. With regards to the accuracy of the measures, proper and systematic method evaluation and validation against gold-standard measures, as was performed in the thesis chapters relating to the inference of sleep, sleep stages and CRF, is required. With regards to confounding, adjustment of models for potential confounding factors will be required. Further, longitudinal studies and causal analyses leveraging genetic instruments could be used in future to ensure the direction of association is correctly modelled. Further, deep learning models applied on multimodal wearable sensor data could make use of Bayesian models. Bayesian models could allow for data description in an interpretable way, something that could help overcome some of the inherent limitations of deep learning, particularly when applied in medicine.

REFERENCES

[1] Mohr, D. C., Zhang, M. & Schueller, S. M. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* **13**, 23–47 (2017).

[2] Rosenman, R., Tennekoon, V. & Hill, L. G. Measuring bias in self-reported data. *International journal of behavioural & healthcare research* **2**, 320–332 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/25383095http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4224297.

[3] Doherty, A. *et al.* Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PLOS ONE* **12**, e0169649 (2017). URL http://dx.plos.org/10.1371/journal.pone.0169649.

[4] German National Cohort (GNC) Consortium. The German National Cohort: aims, study design and organization. *European Journal of Epidemiology* **29**, 371–382 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24840228http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4050302http://link.springer.com/10.1007/s10654-014-9890-7.

[5] Wijndaele, K. *et al.* Utilization and Harmonization of Adult Accelerometry Data: Review and Expert Consensus. *Medicine and science in sports and exercise* **47**, 2129–39 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25785929http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4731236.

[6] MacAuley, D. A history of physical activity, health and medicine. *Journal of the Royal Society of Medicine* **87**, 32 (1994).

[7] Guy, W. A. Contributions to a knowledge of the influence of employments upon health. *Journal of the Statistical Society of London* **6**, 197–211 (1843).

[8] Park, R. J. High-protein diets,"damaged hearts," and rowing men: Antecedents of modern sports medicine and exercise science, 1867–1928. *Exercise and sport sciences reviews* **25**, 137–170 (1997).

[9] Hartley, P. H.-S. & Llewellyn, G. F. Longevity of oarsmen. *British medical journal* **1**, 657 (1939).

[10] Morris, J. N., Heady, J., Raffle, P., Roberts, C. & Parks, J. Coronary heart-disease and physical activity of work. *The Lancet* **262**, 1111–1120 (1953).

[11] Morris, J. N. & Crawford, M. D. Coronary heart disease and physical activity of work. *British medical journal* **2**, 1485 (1958).

[12] Powell, K. E., Thompson, P. D., Caspersen, C. J. & Kendrick, J. S. Physical activity and the incidence of coronary heart. *Ann. Rev* **8**, 253–87 (1987).

References

[13] Pate, R. R. *et al.* Physical activity and public health: a recommendation from the centers for disease control and prevention and the american college of sports medicine. *Jama* **273**, 402–407 (1995).

[14] Calfas, K. J. *et al.* A controlled trial of physician counseling to promote the adoption of physical activity. *Preventive medicine* **25**, 225–233 (1996).

[15] Kahn, E. B. *et al.* The effectiveness of interventions to increase physical activity: a systematic review. *American journal of preventive medicine* **22**, 73–107 (2002).

[16] Troiano, R. P. *et al.* Physical activity in the United States measured by accelerometer. *Medicine and Science in Sports and Exercise* **40**, 181–188 (2008).

[17] Torous, J., Kiang, M. V., Lorme, J. & Onnela, J.-P. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR mental health* **3**, e16 (2016).

[18] Huckvale, K., Venkatesh, S. & Christensen, H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ Digital Medicine* **2**, 1–11 (2019).

[19] Woodward, K., Kanjo, E., Umair, M. & Sas, C. Harnessing digital phenotyping to deliver real-time interventional bio-feedback (2019).

[20] Gordis, L. *Epidemiology*. A student consult title (Elsevier/Saunders, 2009). URL https://books.google.co.uk/books?id=GseHgIbJo4gC.

[21] Wang, J., Li, A. M., Lam, H. S. H. S., Leung, G. M. & Schooling, C. M. Sleep duration and adiposity in children and adults: observational and mendelian randomization studies. *Obesity* **27**, 1013–1022 (2019).

[22] Porta, M. A dictionary of epidemiology. *Revista Española de Salud Pública* **82** (2008).

[23] Froom, P., Melamed, S., Kristal-Boneh, E., Benbassat, J. & Ribak, J. Healthy volunteer effect in industrial workers. *Journal of clinical epidemiology* **52**, 731–735 (1999).

[24] Fry, A. *et al.* Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *American journal of epidemiology* **186**, 1026–1034 (2017).

[25] de Leeuw, E., Borgers, N. & Smits, A. Pretesting Questionnaires for Children and Adolescents. In *Methods for Testing and Evaluating Survey Questionnaires*, 409–429 (John Wiley & Sons, Inc., Hoboken, NJ, USA). URL http://doi.wiley.com/10.1002/0471654728.ch20.

[26] Sallis, J. F. & Saelens, B. E. Assessment of physical activity by self-report: Status, limitations, and future directions. *Research Quarterly for Exercise and Sport* **71**, 1–14 (2000).

[27] Bassett, D. R., Wyatt, H. R., Thompson, H., Peters, J. C. & Hill, J. O. Pedometer-measured physical activity and health behaviors in U.S. adults. *Medicine and Science in Sports and Exercise* **42**, 1819–1825 (2010).

[28] Corder, K., Brage, S. & Ekelund, U. Accelerometers and pedometers: Methodology and clinical application (2007).

[29] Schmidt, M. D., Blizzard, L. C., Venn, A. J., Cochrane, J. A. & Dwyer, T. Practical considerations when using pedometers to assess physical activity in population studies: Lessons from the burnie take heart study. *Research Quarterly for Exercise and Sport* **78**, 162–170 (2007).

[30] Buchman, A. S. *et al.* Total daily physical activity and the risk of AD and cognitive decline in older adults. *Neurology* **78**, 1323–1329 (2012).

[31] Guo, W., Key, T. J. & Reeves, G. K. Accelerometer compared with questionnaire measures of physical activity in relation to body size and composition: A large cross-sectional analysis of UK Biobank. *BMJ Open* **9** (2019).

[32] Willetts, M., Hollowell, S., Aslett, L., Holmes, C. & Doherty, A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Scientific Reports* **8** (2018).

[33] Conn, V. S., Hafdahl, A. R., Brown, S. A. & Brown, L. M. Meta-analysis of patient education interventions to increase physical activity among chronically ill adults (2008).

[34] McCarthy, M. & Grey, M. Motion sensor use for physical activity data: Methodological considerations (2015).

[35] Strath, S. J. *et al.* Guide to the assessment of physical activity: Clinical and research applications: A scientific statement from the American Heart association. *Circulation* **128**, 2259–2279 (2013).

[36] Veltink, P., Bussmann, H., de Vries, W., Martens, W. & Van Lummel, R. Detection of static and dynamic activities using uniaxial accelerometers. *IEEE Transactions on Rehabilitation Engineering* **4**, 375–385 (1996). URL http://ieeexplore.ieee.org/document/547939/.

[37] van Hees, V. T. *et al.* Separating Movement and Gravity Components in an Acceleration Signal and Implications for the Assessment of Human Daily Physical Activity. *PLoS ONE* **8**, e61691 (2013). URL http://dx.plos.org/10.1371/journal.pone.0061691.

[38] Sabatini, A. Quaternion-Based Extended Kalman Filter for Determining Orientation by Inertial and Magnetic Sensing. *IEEE Transactions on Biomedical Engineering* **53**, 1346–1356 (2006). URL http://ieeexplore.ieee.org/document/1643403/.

[39] White, T., Westgate, K., Wareham, N. J. & Brage, S. Estimation of physical activity energy expenditure during free-living from wrist accelerometry in uk adults. *PLoS One* **11** (2016).

[40] Ainsworth, B. E. *et al.* Compendium of physical activities: an update of activity codes and MET intensities. *Medicine and science in sports and exercise* **32**, S498–504 (2000). URL http://www.ncbi.nlm.nih.gov/pubmed/10993420.

[41] ROWLANDS, A. V. *et al.* Sedentary Sphere. *Medicine & Science in Sports & Exercise* **48**, 748–754 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26559451https://insights.ovid.com/crossref?an=00005768-201604000-00022.

[42] Perez-Pozuelo, I. *et al.* Diurnal Profiles of Physical Activity and Postures Derived From Wrist-Worn Accelerometry in UK Adults. *Journal for the Measurement of Physical Behaviour* 1–11 (2019).

[43] Dunstan, D. W., Howard, B., Healy, G. N. & Owen, N. Too much sitting – A health hazard. *Diabetes Research and Clinical Practice* **97**, 368–376 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22682948http://linkinghub.elsevier.com/retrieve/pii/S0168822712002082.

[44] Lambrecht, S. & Del-Ama, A. J. Human movement analysis with inertial sensors. In *Biosystems and Biorobotics*, vol. 4, 305–328 (Springer International Publishing, 2014).

[45] White, R. W., Doraiswamy, P. M. & Horvitz, E. Detecting neurodegenerative disorders from web search signals. *NPJ digital medicine* **1**, 1–4 (2018).

[46] Radu, V. *et al.* Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1**, 1–27 (2018).

[47] Bulling, A., Ward, J. A. & Gellersen, H. Multimodal recognition of reading activity in transit using body-worn sensors. *ACM Transactions on Applied Perception (TAP)* **9**, 2 (2012).

[48] Hemminki, S., Nurmi, P. & Tarkoma, S. Accelerometer-based transportation mode detection on smartphones. In *Proceedings of the 11th ACM conference on embedded networked sensor systems*, 13 (ACM, 2013).

[49] Guo, H., Chen, L., Peng, L. & Chen, G. Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 1112–1123 (ACM, 2016).

[50] Fulcher, B. D. Feature-based time-series analysis. In *Feature Engineering for Machine Learning and Data Analytics*, 87–116 (CRC Press, 2018).

[51] Fulcher, B. D. & Jones, N. S. Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering* **26**, 3026–3037 (2014).

[52] Friedman, J., Hastie, T. & Tibshirani, R. *The elements of statistical learning*, vol. 1 (Springer series in statistics New York, 2001).

[53] Grünerbl, A. *et al.* Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE Journal of Biomedical and Health Informatics* **19**, 140–148 (2015).

[54] Gauss, C. F. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, vol. 7 (Perthes et Besser, 1809).

[55] Bishop, C. M. *Pattern recognition and machine learning* (springer, 2006).

[56] Kandel, E. R. *et al. Principles of neural science*, vol. 4 (McGraw-hill New York, 2000).

[57] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning internal representations by error propagation. Tech. Rep., California Univ San Diego La Jolla Inst for Cognitive Science (1985).

[58] LeCun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. *Neural computation* **1**, 541–551 (1989).

[59] Werbos, P. J. Generalization of backpropagation with application to a recurrent gas market model. *Neural networks* **1**, 339–356 (1988).

[60] Sundermeyer, M., Schlüter, R. & Ney, H. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association* (2012).

[61] Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112 (2014).

[62] Hinton, G. *et al.* Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* **29**, 82–97 (2012).

[63] Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105 (2012).

[64] Oord, A. v. d., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[65] Lee, H.-Y., Huang, J.-B., Singh, M. & Yang, M.-H. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, 667–676 (2017).

[66] Iizuka, S., Simo-Serra, E. & Ishikawa, H. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)* **36**, 107 (2017).

[67] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (2019).

[68] Misra, I., Zitnick, C. L. & Hebert, M. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, 527–544 (Springer, 2016).

[69] Saeed, A., Ozcelebi, T. & Lukkien, J. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **3**, 61 (2019).

[70] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* (2020).

[71] Bulling, A. 33 A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors URL http://dx.doi.org/10.1145/2499621.

[72] Plotz, T. & Guan, Y. Deep Learning for Human Activity Recognition in Mobile Computing. *Computer* **51**, 50–59 (2018). URL https://ieeexplore.ieee.org/document/8364643/.

[73] Bazarevsky, V. *et al.* Blazepose: On-device real-time body pose tracking. *CVPR Workshop on Computer Vision for Augmented and Virtual Reality* (2020).

[74] O'Driscoll, R. *et al.* How well do activity monitors estimate energy expenditure? a systematic review and meta-analysis of the validity of current technologies. *British Journal of Sports Medicine* **54**, 332–340 (2020).

[75] Henriksen, A. *et al.* Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables. *Journal of medical Internet research* **20**, e110 (2018).

[76] Zhai, B., Perez-Pozuelo, I., Clifton, E. A., Palotti, J. & Guan, Y. Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **4**, 1–33 (2020).

[77] Goldsack, J. C. *et al.* Verification, analytical validation, and clinical validation (v3): the foundation of determining fit-for-purpose for biometric monitoring technologies (biomets). *npj digital Medicine* **3**, 1–15 (2020).

[78] Allmark, P. Should research samples reflect the diversity of the population? *Journal of medical ethics* **30**, 185–189 (2004).

[79] Bhopal, R. S. *Ethnicity, race, and health in multicultural societies: foundations for better epidemiology, public health, and health care* (Oxford University Press, 2007).

[80] Bent, B., Goldstein, B. A., Kibbe, W. A. & Dunn, J. P. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine* **3**, 1–9 (2020).

[81] Sadeh, A., Sharkey, K. M. & Carskadon, M. A. Activity-Based Sleep—Wake Identification: An Empirical Test of Methodological Issues. *Sleep* **17**, 201–207 (1994).

[82] Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J. & Gillin, J. C. Automatic sleep/wake identification from wrist activity. *Sleep* **15**, 461–9 (1992). URL http://www.ncbi.nlm.nih.gov/pubmed/1455130.

[83] Kripke, D. F. *et al.* Wrist actigraphic scoring for sleep laboratory patients: algorithm development. *Journal of sleep research* **19**, 612–619 (2010).

[84] Sazonov, E., Sazonova, N., Schuckers, S., Neuman, M. & CHIME Study Group. Activity-based sleep-wake identification in infants. *Physiological Measurement* **25**, 1291–1304 (2004).

[85] Arora, T., Broglia, E., Pushpakumar, D., Lodhi, T. & Taheri, S. An Investigation into the Strength of the Association and Agreement Levels between Subjective and Objective Sleep Duration in Adolescents. *PLOS ONE* **8**, e72406 (2013). URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0072406. Publisher: Public Library of Science.

[86] Aili Katarina, Åström Paulsson Sofia, Stoetzer Ulrich, Svartengren Magnus & Hillert Lena. Reliability of Actigraphy and Subjective Sleep Measurements in Adults: The Design of Sleep Assessments. *Journal of Clinical Sleep Medicine* **13**, 39–47 (2017). URL https://jcsm.aasm.org/doi/10.5664/jcsm.6384. Publisher: American Academy of Sleep Medicine.

[87] Park, H. & Suh, B. Association between sleep quality and physical activity according to gender and shift work. *Journal of Sleep Research* **n/a**, e12924 (2019). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.12924. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.12924.

[88] Katzmarzyk, P. T., Church, T. S., Janssen, I., Ross, R. & Blair, S. N. Metabolic syndrome, obesity, and mortality: impact of cardiorespiratory fitness. *Diabetes care* **28**, 391–397 (2005).

[89] Wei, M. *et al.* The association between cardiorespiratory fitness and impaired fasting glucose and type 2 diabetes mellitus in men. *Annals of internal medicine* **130**, 89–96 (1999).

[90] Schmid, D. & Leitzmann, M. Cardiorespiratory fitness as predictor of cancer mortality: a systematic review and meta-analysis. *Annals of oncology* **26**, 272–278 (2015).

[91] Blair, S. N. *et al.* Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women. *Jama* **276**, 205–210 (1996).

[92] Kodama, S. *et al.* Cardiorespiratory fitness as a quantitative predictor of all-cause mortality and cardiovascular events in healthy men and women: a meta-analysis. *Jama* **301**, 2024–2035 (2009).

[93] Jensen, M. T., Marott, J. L. & Jensen, G. B. Elevated resting heart rate is associated with greater risk of cardiovascular and all-cause mortality in current and former smokers. *International journal of cardiology* **151**, 148–154 (2011).

[94] Jensen, M. T., Suadicani, P., Hein, H. O. & Gyntelberg, F. Elevated resting heart rate, physical fitness and all-cause mortality: a 16-year follow-up in the copenhagen male study. *Heart* **99**, 882–887 (2013).

[95] Cooney, M. T. *et al.* Elevated resting heart rate is an independent risk factor for cardiovascular disease in healthy men and women. *American heart journal* **159**, 612–619 (2010).

[96] Zhang, D., Wang, W. & Li, F. Association between resting heart rate and coronary artery disease, stroke, sudden death and noncardiovascular diseases: a meta-analysis. *Cmaj* **188**, E384–E392 (2016).

[97] Lee, D. H. *et al.* Resting heart rate as a prognostic factor for mortality in patients with breast cancer. *Breast cancer research and treatment* **159**, 375–384 (2016).

[98] Seviiri, M. *et al.* Resting heart rate, temporal changes in resting heart rate, and overall and cause-specific mortality. *Heart* **104**, 1076–1085 (2018).

[99] Gonzales, T. I. *et al.* Resting heart rate as a biomarker for tracking change in cardiorespiratory fitness of uk adults: The fenland study. *medRxiv* (2020).

[100] Daumé III, H. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815* (2009).

[101] Jiang, J. & Zhai, C. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, 264–271 (2007).

[102] Long, M., Zhu, H., Wang, J. & Jordan, M. I. Unsupervised domain adaptation with residual transfer networks. In *Advances in neural information processing systems*, 136–144 (2016).

[103] Wang, J., Chen, Y., Hu, L., Peng, X. & Philip, S. Y. Stratified transfer learning for cross-domain activity recognition. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 1–10 (IEEE, 2018).

[104] Akbari, A. & Jafari, R. Transferring activity recognition models for new wearable sensors with deep generative domain adaptation. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*, 85–96 (2019).

[105] Rokni, S. A. & Ghasemzadeh, H. Synchronous dynamic view learning: a framework for autonomous training of activity recognition models using wearable sensors. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*, 79–90 (2017).

[106] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two-sample test. *Journal of Machine Learning Research* **13**, 723–773 (2012).

[107] Long, M., Cao, Y., Wang, J. & Jordan, M. I. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791* (2015).

[108] Ganin, Y. & Lempitsky, V. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495* (2014).

[109] Ayas, N. T. *et al.* A prospective study of self-reported sleep duration and incident diabetes in women. *Diabetes care* **26**, 380–384 (2003).

[110] Reid, K. J. *et al.* Aerobic exercise improves self-reported sleep and quality of life in older adults with insomnia. *Sleep medicine* **11**, 934–940 (2010).

[111] Yuda, E. *et al.* Sleep stage classification by a combination of actigraphic and heart rate signals. *Journal of Low Power Electronics and Applications* **7**, 28 (2017).

[112] Doherty, A. *et al.* Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study. *PloS one* **12** (2017).

[113] Perez-Pozuelo, I. *et al.* Diurnal profiles of physical activity and postures derived from wrist-worn accelerometry in uk adults. *Journal for the Measurement of Physical Behaviour* **1**, 1–11 (2019).

[114] Hartley, P. *et al.* Using accelerometers to measure physical activity in older patients admitted to hospital. *Current Gerontology and Geriatrics Research* **2018**.

[115] Moreira, J. B. N., Wohlwend, M. & Wisløff, U. Exercise and cardiac health: physiological and molecular insights. *Nature Metabolism* (2020). URL https://doi.org/10.1038/s42255-020-0262-1.

[116] Radu, V. *et al.* Towards multimodal deep learning for activity recognition on mobile devices. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 185–188 (2016).

[117] Ngiam, J. *et al.* Multimodal deep learning. In *ICML* (2011).

[118] Bonomi, L., Huang, Y. & Ohno-Machado, L. Privacy challenges and research opportunities for genomic data sharing. *Nature Genetics* 1–9 (2020).

[119] Schwartz, J. R. L. & Roth, T. Neurophysiology of sleep and wakefulness: basic science and clinical implications. *Current neuropharmacology* **6**, 367–78 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/19587857http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2701283.

[120] Imeri, L. & Opp, M. R. How (and why) the immune system makes us sleep. *Nature reviews. Neuroscience* **10**, 199–210 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19209176http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2839418.

[121] Dawson, D. & Reid, K. Fatigue, alcohol and performance impairment. *Nature* **388**, 235 (1997).

[122] Bertisch, S. M. *et al.* Insomnia with objective short sleep duration and risk of incident cardiovascular disease and all-cause mortality: Sleep heart health study. *Sleep* **41**, zsy047 (2018).

[123] Bonnet, M. H. & Arand, D. L. We are chronically sleep deprived. *Sleep* **18**, 908–911 (1995).

[124] Drake, C. L., Roehrs, T., Richardson, G., Walsh, J. K. & Roth, T. Shift work sleep disorder: prevalence and consequences beyond that of symptomatic day workers. *Sleep* **27**, 1453–1462 (2004).

[125] Dement, W. C. & Vaughan, C. C. *The promise of sleep: A pioneer in sleep medicine explores the vital connection between health, happiness, and a good night's sleep* (Delacorte Press New York, 1999).

[126] Groeger, J. A., Zijlstra, F. & Dijk, D.-J. Sleep quantity, sleep difficulties and their perceived consequences in a representative sample of some 2000 british adults. *Journal of sleep research* **13**, 359–371 (2004).

[127] Hafner, M., Stepanek, M., Taylor, J., Troxel, W. M. & van Stolk, C. Why sleep matters—the economic costs of insufficient sleep: a cross-country comparative analysis. *Rand health quarterly* **6** (2017).

[128] Hillman, D. R., Murphy, A. S., Antic, R. & Pezzullo, L. The economic cost of sleep disorders. *Sleep* **29**, 299–305 (2006).

[129] Ozminkowski, R. J., Wang, S. & Walsh, J. K. The direct and indirect costs of untreated insomnia in adults in the united states. *Sleep* **30**, 263–273 (2007).

[130] Ohayon, M. *et al.* National sleep foundation's sleep quality recommendations: first report. *Sleep Health* **3**, 6–19 (2017).

[131] Taheri, S. The link between short sleep duration and obesity: we should recommend more sleep to prevent obesity. *Archives of Disease in Childhood* **91**, 881–884 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/17056861http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2082964http://adc.bmj.com/cgi/doi/10.1136/adc.2005.093013.

[132] Awad, K. M., Malhotra, A., Barnet, J. H., Quan, S. F. & Peppard, P. E. Exercise is associated with a reduced incidence of sleep-disordered breathing. *The American journal of medicine* **125**, 485–490 (2012).

[133] St-Onge, M.-P., Mikic, A. & Pietrolungo, C. E. Effects of diet on sleep quality. *Advances in Nutrition* **7**, 938–949 (2016).

[134] Kline, C. E. The bidirectional relationship between exercise and sleep: implications for exercise adherence and sleep improvement. *American journal of lifestyle medicine* **8**, 375–379 (2014).

[135] Walker, M. *Why we sleep: The new science of sleep and dreams* (Penguin UK, 2017).

[136] Shan, Z. *et al.* Sleep duration and risk of type 2 diabetes: a meta-analysis of prospective studies. *Diabetes care* **38**, 529–537 (2015).

[137] Wulff, K., Gatti, S., Wettstein, J. G. & Foster, R. G. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nature Reviews Neuroscience* **11**, 589 (2010).

[138] Marshall, N. S. *et al.* Sleep apnea as an independent risk factor for all-cause mortality: the busselton health study. *Sleep* **31**, 1079–1085 (2008).

[139] Cappuccio, F. P., Cooper, D., D'elia, L., Strazzullo, P. & Miller, M. A. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. *European heart journal* **32**, 1484–1492 (2011).

[140] King, C. R. *et al.* Short sleep duration and incident coronary artery calcification. *Jama* **300**, 2859–2866 (2008).

[141] Chandola, T., Ferrie, J. E., Perski, A., Akbaraly, T. & Marmot, M. G. The effect of short sleep duration on coronary heart disease risk is greatest among those with sleep disturbance: a prospective study from the whitehall ii cohort. *Sleep* **33**, 739–744 (2010).

[142] Nagai, M., Hoshide, S. & Kario, K. Sleep duration as a risk factor for cardiovascular disease-a review of the recent literature. *Current cardiology reviews* **6**, 54–61 (2010).

[143] Lin, X. *et al.* Night-shift work increases morbidity of breast cancer and all-cause mortality: a meta-analysis of 16 prospective cohort studies. *Sleep medicine* **16**, 1381–1387 (2015).

[144] Knutson, K. L., Spiegel, K., Penev, P. & Van Cauter, E. The metabolic consequences of sleep deprivation. *Sleep medicine reviews* **11**, 163–178 (2007).

[145] Ju, Y.-E. S., Lucey, B. P. & Holtzman, D. M. Sleep and Alzheimer disease pathology–a bidirectional relationship. *Nature reviews. Neurology* **10**, 115–9 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24366271http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3979317.

[146] Spira, A. P., Chen-Edinboro, L. P., Wu, M. N. & Yaffe, K. Impact of sleep on the risk of cognitive decline and dementia. *Current opinion in psychiatry* **27**, 478 (2014).

[147] Brown, B. M., Rainey-Smith, S. R., Bucks, R. S., Weinborn, M. & Martins, R. N. Exploring the bi-directional relationship between sleep and beta-amyloid. *Current opinion in psychiatry* **29**, 397–401 (2016).

[148] Becker, N. B. *et al.* Depression and quality of life in older adults: Mediation effect of sleep quality. *International Journal of Clinical and Health Psychology* **18**, 8–17 (2018).

[149] Besedovsky, L., Lange, T. & Born, J. Sleep and immune function. *Pflügers Archiv-European Journal of Physiology* **463**, 121–137 (2012).

[150] Lu, Y., Tian, N., Yin, J., Shi, Y. & Huang, Z. Association between sleep duration and cancer risk: a meta-analysis of prospective cohort studies. *PloS one* **8**, e74723 (2013).

[151] Blask, D. E. Melatonin, sleep disturbance and cancer risk. *Sleep medicine reviews* **13**, 257–264 (2009).

[152] Spiegel, K., Sheridan, J. F. & Van Cauter, E. Effect of sleep deprivation on response to immunizaton. *Jama* **288**, 1471–1472 (2002).

[153] Jaiswal, S. J., Topol, E. J. & Steinhubl, S. R. Digitising the way to better sleep health. *The Lancet* **393**, 639 (2019).

[154] Morgenthaler, T. I. *et al.* Practice parameters for the clinical evaluation and treatment of circadian rhythm sleep disorders. An American Academy of Sleep Medicine report. *Sleep* **30**, 1445–59 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/18041479http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2082098.

[155] Hao, Y. & Foster, R. Wireless body sensor networks for health-monitoring applications. *Physiological Measurement* **29**, R27–R56 (2008).

[156] Shepard, J. W. *et al.* History of the development of sleep medicine in the united states. *Journal of clinical sleep medicine* **1**, 61–82 (2005).

[157] Phelps, A. J. *et al.* An Ambulatory Polysomnography Study of the Post-traumatic Nightmares of Post-traumatic Stress Disorder. *SLEEP* **41** (2018). URL http://dx.doi.org/10.1093/sleep/zsx188.

[158] Schwichtenberg, A. J., Choe, J., Kellerman, A., Abel, E. A. & Delp, E. J. Pediatric Videosomnography: Can Signal/Video Processing Distinguish Sleep and Wake States? *Frontiers in pediatrics* **6**, 158 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29974042http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6020776.

[159] Aggarwal, K., Khadanga, S., Joty, S., Kazaglis, L. & Srivastava, J. A structured learning approach with neural conditional random fields for sleep staging. In *2018 IEEE International Conference on Big Data (Big Data)*, 1318–1327 (IEEE, 2018).

[160] Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**, 423–443 (2018).

[161] Caulfield, B., Reginatto, B. & Slevin, P. Not all sensors are created equal: a framework for evaluating human performance measurement technologies. *npj Digital Medicine* **2**, 7 (2019). URL http://www.nature.com/articles/s41746-019-0082-4.

[162] Troiano, R. P. *et al.* Physical activity in the united states measured by accelerometer. *Medicine & Science in Sports & Exercise* **40**, 181–188 (2008).

[163] Sadeh, A. The role and validity of actigraphy in sleep medicine: an update. *Sleep medicine reviews* **15**, 259–267 (2011).

[164] Martin, J. L. & Hakim, A. D. Wrist actigraphy. *Chest* **139**, 1514–1527 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21652563http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3109647.

[165] Moon, Y. *et al.* Monitoring gait in multiple sclerosis with novel wearable motion sensors. *PLOS ONE* **12**, e0171346 (2017).

[166] Tal, A., Shinar, Z., Shaki, D., Codish, S. & Goldbart, A. Validation of Contact-Free Sleep Monitoring Device with Comparison to Polysomnography. *Journal of Clinical Sleep Medicine* **13**, 517–522 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/27998378http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5337599http://jcsm.aasm.org/ViewAbstract.aspx?pid=30976.

[167] Paalasmaa, J., Leppakorpi, L. & Partinen, M. Quantifying respiratory variation with force sensor measurements. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2011, 3812–3815 (IEEE, 2011). URL http://www.ncbi.nlm.nih.gov/pubmed/22255170http://ieeexplore.ieee.org/document/6090773/.

[168] Paalasmaa, J., Toivonen, H. & Partinen, M. Adaptive Heartbeat Modeling for Beat-to-Beat Heart Rate Measurement in Ballistocardiograms. *IEEE Journal of Biomedical and Health Informatics* **19**, 1945–1952 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/24691540http://ieeexplore.ieee.org/document/6780577/.

[169] Chow, P., Nagendra, G., Abisheganaden, J. & Wang, Y. Respiratory monitoring using an air-mattress system. *Physiological measurement* **21**, 345 (2000).

[170] Chee, Y., Han, J., Youn, J. & Park, K. Air mattress sensor system with balancing tube for unconstrained measurement of respiration and heart beat movements. *Physiological measurement* **26**, 413 (2005).

[171] Arlotto, P., Grimaldi, M., Naeck, R. & Ginoux, J.-M. An ultrasonic contactless sensor for breathing monitoring. *Sensors* **14**, 15371–15386 (2014).

[172] Sadek, I., Bellmunt, J., Kodyš, M., Abdulrazak, B. & Mokhtari, M. Novel unobtrusive approach for sleep monitoring using fiber optics in an ambient assisted living platform. In *International Conference on Smart Homes and Health Telematics*, 48–60 (Springer, 2017).

[173] Chen, Z. *et al.* Simultaneous measurement of breathing rate and heart rate using a microbend multimode fiber optic sensor. *Journal of biomedical optics* **19**, 057001 (2014).

[174] Kam, J. W. *et al.* Systematic comparison between a wireless eeg system with dry electrodes and a wired eeg system with wet electrodes. *NeuroImage* **184**, 119–129 (2019).

[175] Finan, P. H. *et al.* Validation of a Wireless, Self-Application, Ambulatory Electroencephalographic Sleep Monitoring Device in Healthy Volunteers. *Journal of Clinical Sleep Medicine* **12**, 1443–1451 (2016). URL http://jcsm.aasm.org/ViewAbstract.aspx?pid=30850.

[176] Koley, B. & Dey, D. An ensemble system for automatic sleep stage classification using single channel EEG signal. *Computers in Biology and Medicine* **42**, 1186–1195 (2012). URL https://www.sciencedirect.com/science/article/pii/S0010482512001588.

[177] Myllymaa, S. *et al.* Assessment of the suitability of using a forehead eeg electrode set and chin emg electrodes for sleep staging in polysomnography. *Journal of sleep research* **25**, 636–645 (2016).

[178] Looney, D., Goverdovsky, V., Rosenzweig, I., Morrell, M. J. & Mandic, D. P. Wearable in-ear encephalography sensor for monitoring sleep. preliminary observations from nap studies. *Annals of the American Thoracic Society* **13**, 2229–2233 (2016).

[179] Mikkelsen, K. B. *et al.* Machine-learning-derived sleep–wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy. *Journal of sleep research* e12786 (2018).

[180] Nakamura, T., Alqurashi, Y. D., Morrell, M. J. & Mandic, D. Hearables: automatic overnight sleep monitoring with standardised in-ear eeg sensor. *IEEE Transactions on Biomedical Engineering* (2019).

[181] Wang, F., Li, G., Chen, J., Duan, Y. & Zhang, D. Novel semi-dry electrodes for brain–computer interface applications. *Journal of neural engineering* **13**, 046021 (2016).

[182] Borger, J. N., Huber, R. & Ghosh, A. Capturing sleep–wake cycles by using day-to-day smartphone touchscreen interactions. *npj Digital Medicine* **2**, 73 (2019). URL http://www.nature.com/articles/s41746-019-0147-4.

[183] Hao, T., Xing, G. & Zhou, G. iSleep: unobtrusive sleep quality monitoring using smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, 4 (ACM, 2013).

[184] Ong, A. A. & Gillespie, M. B. Overview of smartphone applications for sleep analysis. *World Journal of Otorhinolaryngology - Head and Neck Surgery* **2**, 45 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/29204548http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5698521.

[185] Min, S. D., Yoon, D. J., Yoon, S. W., Yun, Y. H. & Lee, M. A study on a non-contacting respiration signal monitoring system using doppler ultrasound. *Medical & biological engineering & computing* **45**, 1113–1119 (2007).

[186] Shahshahani, A., Bhadra, S. & Zilic, Z. A continuous respiratory monitoring system using ultrasound piezo transducer. In *Circuits and Systems (ISCAS), 2018 IEEE International Symposium on*, 1–4 (IEEE, 2018).

[187] Rahman, T. *et al.* DoppleSleep: A Contactless Unobtrusive Sleep Sensing System Using Short-Range Doppler Radar. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 39–50 (ACM, 2015). URL http://dl.acm.org/citation.cfm?doid=2750858.2804280.

[188] Nijsure, Y. *et al.* An impulse radio ultrawideband system for contactless noninvasive respiratory monitoring. *IEEE Trans. Biomed. Engineering* **60**, 1509–1517 (2013).

[189] Kaltiokallio, O. J., Yigitler, H., Jäntti, R. & Patwari, N. Non-invasive respiration rate monitoring using a single cots tx-rx pair. In *Proceedings of the 13th international symposium on Information processing in sensor networks*, 59–70 (IEEE Press, 2014).

[190] Adib, F., Mao, H., Kabelac, Z., Katabi, D. & Miller, R. C. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 837–846 (ACM, 2015).

[191] Droitcour, A. D., Boric-Lubecke, O. & Kovacs, G. T. Signal-to-noise ratio in doppler radar system for heart and respiratory rate measurements. *IEEE transactions on microwave theory and techniques* **57**, 2498–2507 (2009).

[192] Zhao, M., Yue, S., Katabi, D., Jaakkola, T. S. & Bianchi, M. T. Learning Sleep Stages from Radio Signals: A Conditional Adversarial Architecture. In *Proceedings of the 34th International Conference on Machine Learning*, 4100–4109 (2017). URL http://sleep.csail.mit.edu/.

[193] Hsu, C.-Y. *et al.* Zero-effort in-home sleep and insomnia monitoring using radio signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1**, 59 (2017).

[194] Tataraidze, A. *et al.* Bioradiolocation-based sleep stage classification. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, 2839–2842 (IEEE, 2016).

[195] Nam, Y., Kim, Y. & Lee, J. Sleep monitoring based on a tri-axial accelerometer and a pressure sensor. *Sensors* **16**, 750 (2016).

[196] Radha, M. *et al.* Lstm knowledge transfer for hrv-based sleep staging. *arXiv preprint arXiv:1809.06221* (2018).

[197] Yasumoto, K., Yamaguchi, H. & Shigeno, H. Survey of Real-time Processing Technologies of IoT Data Streams. *Journal of Information Processing* **24**, 195–202 (2016). URL https://www.jstage.jst.go.jp/article/ipsjjip/24/2/24_195/_article/-char/ja/.

[198] Bragazzi, N. L., Guglielmi, O. & Garbarino, S. SleepOMICS: How Big Data Can Revolutionize Sleep Science. *International Journal of Environmental Research and Public Health* **16** (2019). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351921/.

[199] Yacchirema, D. C., Sarabia-Jácome, D., Palau, C. E. & Esteve, M. A Smart System for Sleep Monitoring by Integrating IoT With Big Data Analytics. *IEEE Access* **6**, 35988–36001 (2018).

[200] Chiang, M. & Zhang, T. Fog and iot: An overview of research opportunities. *IEEE Internet of Things Journal* **3**, 854–864 (2016).

[201] Yousefpour, A. *et al.* All one needs to know about fog computing and related edge computing paradigms: A complete survey. *Journal of Systems Architecture* (2019).

[202] Aazam, M. & Huh, E.-N. Fog computing and smart gateway based communication for cloud of things. In *2014 International Conference on Future Internet of Things and Cloud*, 464–470 (IEEE, 2014).

[203] Hsieh, Y.-Z. Internet of things pillow detecting sleeping quality. In *2018 1st International Cognitive Cities Conference (IC3)*, 266–267 (IEEE, 2018).

[204] Sangat, P., Indrawan-Santiago, M. & Taniar, D. Sensor data management in the cloud: Data storage, data ingestion, and data retrieval. *Concurrency and Computation: Practice and Experience* **30**, e4354 (2018). URL https://onlinelibrary.wiley.com/doi/full/10.1002/cpe.4354.

[205] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S. & Stoica, I. Spark: Cluster computing with working sets. *HotCloud* **10**, 95 (2010).

[206] Van Drongelen, W. *Signal processing for neuroscientists* (Academic press, 2018).

[207] Devasahayam, S. R. *Signals and systems in biomedical engineering: signal processing and physiological systems modeling* (Springer Science & Business Media, 2012).

[208] Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* **25**, 44 (2019).

[209] Ramesh, A., Kambhampati, C., Monson, J. R. & Drew, P. Artificial intelligence in medicine. *Annals of The Royal College of Surgeons of England* **86**, 334 (2004).

[210] Shahin, M. *et al.* Deep learning and insomnia: Assisting clinicians with their diagnosis. *IEEE journal of biomedical and health informatics* **21**, 1546–1553 (2017).

[211] Malafeev, A. *et al.* Automatic Human Sleep Stage Scoring Using Deep Neural Networks. *Frontiers in Neuroscience* **12**, 781 (2018). URL https://www.frontiersin.org/article/10.3389/fnins.2018.00781/full.

[212] Bauer, J. S. *et al.* ShutEye: Encouraging Awareness of Healthy Sleep Recommendations with a Mobile, Peripheral Display. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, 1401–1410. ACM (ACM Press, 2012). URL http://dl.acm.org/citation.cfm?doid=2207676.2208600.

[213] Choi, Y. K. *et al.* Smartphone applications to support sleep self-management: review and evaluation. *Journal of Clinical Sleep Medicine* **14**, 1783–1790 (2018).

[214] Bhat, S. *et al.* Is There a Clinical Role For Smartphone Sleep Apps? Comparison of Sleep Cycle Detection by a Smartphone Application to Polysomnography. *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine* **11**, 709–15 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25766719http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4481053.

[215] Majumder, S. *et al.* Smart Homes for Elderly Healthcare-Recent Advances and Research Challenges. *Sensors (Basel, Switzerland)* **17** (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/29088123http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5712846.

[216] Sateia, M. J. International classification of sleep disorders. *Chest* **146**, 1387–1394 (2014).

[217] Subramanian, S., Hesselbacher, S., Mattewal, A. & Surani, S. Gender and age influence the effects of slow-wave sleep on respiration in patients with obstructive sleep apnea. *Sleep and Breathing* **17**, 51–56 (2013).

[218] Rosenberg, R. S. & Van Hout, S. The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring. *Journal of clinical sleep medicine* **9**, 81–87 (2013).

[219] Danker-hopfe, H. *et al.* Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard. *Journal of sleep research* **18**, 74–84 (2009).

[220] Biswal, S. *et al.* Expert-level sleep scoring with deep neural networks. *Journal of the American Medical Informatics Association* **25**, 1643–1650 (2018).

[221] Sadeh, A., Sharkey, K. M. & Carskadon, M. A. Activity-based sleep-wake identification: an empirical test of methodological issues. *Sleep* **17**, 201–207 (1994).

[222] Penzel, T. *et al.* Digital analysis and technical specifications. *Journal of clinical sleep medicine* **3**, 109–120 (2007).

[223] Palotti, J. *et al.* Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *npj Digital Medicine* **2**, 50 (2019).

[224] Yan, R. *et al.* Multi-modality of polysomnography signals' fusion for automatic sleep scoring. *Biomedical Signal Processing and Control* **49**, 14–23 (2019). URL https://www.sciencedirect.com/science/article/pii/S1746809418302647.

[225] Sano, A. & Picard, R. W. Recognition of sleep dependent memory consolidation with multi-modal sensor data. In *2013 IEEE International Conference on Body Sensor Networks*, 1–4 (IEEE, 2013). URL http://ieeexplore.ieee.org/document/6575479/.

[226] LeCun, Y., Bengio, Y. *et al.* Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **3361**, 1995 (1995).

[227] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).

[228] Aggarwal, K., Joty, S., Fernandez-Luque, L. & Srivastava, J. Adversarial unsupervised representation learning for activity time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 834–841 (2019).

[229] Zhang, Y. *et al.* A Comparison Study on Multidomain EEG Features for Sleep Stage Classification. *Computational intelligence and neuroscience* **2017**, 4574079 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/29230239http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5694609.

[230] Giannakeas Biomed Sci, N. J. & Res, T. EEG-Based Automatic Sleep Stage Classification. *Biomedical Journal of Scientific & Technical Research* **5** (2018). URL https://biomedres.us/pdfs/BJSTR.MS.ID.001535.pdf.

[231] Park, J., Kim, D., Yang, C. & Ko, H. SVM based dynamic classifier for sleep disorder monitoring wearable device. In *2016 IEEE International Conference on Consumer Electronics (ICCE)*, 309–310 (IEEE, 2016). URL http://ieeexplore.ieee.org/document/7430624/.

[232] Pan, S.-T., Kuo, C.-E., Zeng, J.-H. & Liang, S.-F. A transition-constrained discrete hidden Markov model for automatic sleep staging. *Biomedical engineering online* **11**, 52 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22908930http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3462123.

[233] Huang, Q. *et al.* Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data. *Journal of the Royal Society, Interface* **15** (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29436510http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5832732.

[234] Yılmaz, B., Asyalı, M. H., Arıkan, E., Yetkin, S. & Özgen, F. Sleep stage and obstructive apneaic epoch classification using single-lead ecg. *Biomedical engineering online* **9**, 39 (2010).

[235] Khalighi, S., Sousa, T. & Nunes, U. Adaptive automatic sleep stage classification under covariate shift. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2012, 2259–2262 (IEEE, 2012). URL http://www.ncbi.nlm.nih.gov/pubmed/23366373http://ieeexplore.ieee.org/document/6346412/.

[236] Fonseca, P., den Teuling, N., Long, X. & Aarts, R. M. A comparison of probabilistic classifiers for sleep stage classification. *Physiological Measurement* **39**, 055001 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29620019http://stacks.iop.org/0967-3334/39/i=5/a=055001?key=crossref.a13d909830f251d04926f7eb75c7269b.

[237] Palotti, J. *et al.* Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *npj Digital Medicine* **2**, 50 (2019). URL http://www.nature.com/articles/s41746-019-0126-9.

[238] Lajnef, T. *et al.* Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. *Journal of Neuroscience Methods* **250**, 94–105 (2015). URL https://www.sciencedirect.com/science/article/pii/S0165027015000230.

[239] Samy, L., Huang, M.-C., Liu, J. J., Xu, W. & Sarrafzadeh, M. Unobtrusive sleep stage identification using a pressure-sensitive bed sheet. *IEEE Sensors Journal* **14**, 2092–2101 (2013).

[240] Hassan, A. R. & Bhuiyan, M. I. H. Automatic sleep scoring using statistical features in the emd domain and ensemble methods. *Biocybernetics and Biomedical Engineering* **36**, 248–255 (2016).

[241] Hassan, A. R., Bashar, S. K. & Bhuiyan, M. I. H. On the classification of sleep states by means of statistical and spectral features from single channel electroencephalogram. In *2015 International conference on advances in computing, communications and informatics (ICACCI)*, 2238–2243 (IEEE, 2015).

[242] Radha, M., Garcia-Molina, G., Poel, M. & Tononi, G. Comparison of feature and classifier algorithms for online automatic sleep staging based on a single eeg signal. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1876–1880 (IEEE, 2014).

[243] Reimer, U., Emmenegger, S., Maier, E., Zhang, Z. & Khatami, R. Recognizing sleep stages with wearable sensors in everyday settings. In *ICT4AgeingWell*, 172–179 (2017).

[244] Pouyan, M. B., Nourani, M. & Pompeo, M. Sleep state classification using pressure sensor mats. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1207–1210 (IEEE, 2015).

[245] Wang, Q., Zhao, D., Wang, Y. & Hou, X. Ensemble learning algorithm based on multi-parameters for sleep staging. *Medical & biological engineering & computing* 1–15 (2019).

[246] Na, S., Xumin, L. & Yong, G. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on intelligent information technology and security informatics*, 63–67 (IEEE, 2010).

[247] Acharya, U. R., Chua, E. C.-P., Chua, K. C., Min, L. C. & Tamura, T. Analysis and automatic identification of sleep stages using higher order spectra. *International journal of neural systems* **20**, 509–521 (2010).

[248] Tsinalis, O., Matthews, P. M., Guo, Y. & Zafeiriou, S. Automatic sleep stage scoring with single-channel eeg using convolutional neural networks. *arXiv preprint arXiv:1610.01683* (2016).

[249] Biswal, S. *et al.* Sleepnet: automated sleep staging system via deep learning. *arXiv preprint arXiv:1707.08262* (2017). URL http://arxiv.org/abs/1707.08262.

[250] Zhang, X. *et al.* Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device. *Computers in biology and medicine* **103**, 71–81 (2018).

[251] Chen, W. *et al.* Multimodal ambulatory sleep detection. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 465–468 (IEEE, 2017).

[252] Dursun, M., Gunes, S., Ozsen, S. & Yosunkaya, S. Comparison of artificial immune clustering with fuzzy c-means clustering in the sleep stage classification problem. In *2012 International Symposium on Innovations in Intelligent Systems and Applications*, 1–4 (IEEE, 2012).

[253] Correa, A. G. & Leber, E. L. An automatic detector of drowsiness based on spectral analysis and wavelet decomposition of eeg records. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 1405–1408 (IEEE, 2010).

[254] Oakley, N. Validation with polysomnography of the sleepwatch sleep/wake scoring algorithm used by the actiwatch activity monitoring system. *Bend: Mini Mitter, Cambridge Neurotechnology* (1997).

[255] Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J. & Gillin, J. C. Automatic sleep/wake identification from wrist activity. *Sleep* **15**, 461–469 (1992).

[256] Webster, J. B., Kripke, D. F., Messin, S., Mullaney, D. J. & Wyborney, G. An activity-based sleep monitor system for ambulatory use. *Sleep* **5**, 389–399 (1982).

[257] Jean-Louis, G. *et al.* Determination of sleep and wakefulness with the actigraph data analysis software (adas). *Sleep* **19**, 739–743 (1996).

[258] Chen, Z. & Liu, B. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **10**, 1–145 (2016).

[259] Finelli, L. A., Achermann, P. & Borbély, A. A. Individual 'fingerprints' in human sleep eeg topography. *Neuropsychopharmacology* **25**, S57 (2001).

[260] Buckelmüller, J., Landolt, H.-P., Stassen, H. & Achermann, P. Trait-like individual differences in the human sleep electroencephalogram. *Neuroscience* **138**, 351–356 (2006).

[261] Mikkelsen, K. & de Vos, M. Personalizing deep learning models for automatic sleep staging. *arXiv preprint arXiv:1801.02645* (2018).

[262] Yin, Z., Wang, Y., Liu, L., Zhang, W. & Zhang, J. Cross-subject eeg feature selection for emotion recognition using transfer recursive feature elimination. *Frontiers in neurorobotics* **11**, 19 (2017).

[263] Jiang, Y. *et al.* Seizure classification from eeg signals using transfer learning, semi-supervised learning and tsk fuzzy system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **25**, 2270–2284 (2017).

[264] Konečnỳ, J. *et al.* Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).

[265] Guan, Y., Li, C. & Roli, F. On reducing the effect of covariate factors in gait recognition: A classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 1521–1528 (2015).

[266] Pillay, K. *et al.* Automated eeg sleep staging in the term-age baby using a generative modelling approach. *Journal of neural engineering* **15**, 036004 (2018).

[267] Smith, M. T. *et al.* Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an american academy of sleep medicine clinical practice guideline. *Journal of Clinical Sleep Medicine* **14**, 1231–1237 (2018).

[268] Stretch, R. *et al.* Sleepdb: A clinical and administrative database developed to improve the diagnosis, management and longitudinal tracking of sleep disorders. In *A34. SCREENING, DIAGNOSIS, AND TREATMENT IN SLEEP DISORDERS*, A1389–A1389 (American Thoracic Society, 2019).

[269] Stenholm, S. *et al.* Sleep duration and sleep disturbances as predictors of healthy and chronic disease–free life expectancy between ages 50 and 75: A pooled analysis of three cohorts. *The Journals of Gerontology: Series A* **74**, 204–210 (2018).

[270] Castell, M., Makovski, T., Bocquet, V. & Stranges, S. Sleep duration and multimorbidity in luxembourg. results from the european health examination survey. *Revue d'Épidémiologie et de Santé Publique* **66**, S414 (2018).

[271] Fox, R. S. *et al.* Sleep disturbance and cancer-related fatigue symptom cluster in breast cancer patients undergoing chemotherapy. *Supportive Care in Cancer* 1–11 (2019).

[272] Jung, D. *et al.* Longitudinal association of poor sleep quality with chemotherapy-induced nausea and vomiting in patients with breast cancer. *Psychosomatic medicine* **78**, 959–965 (2016).

[273] Braley, T. J., Kratz, A. L., Kaplish, N. & Chervin, R. D. Sleep and cognitive function in multiple sclerosis. *Sleep* **39**, 1525–1533 (2016).

[274] Ashare, R. L. *et al.* Sleep disturbance during smoking cessation: Withdrawal or side effect of treatment? *Journal of smoking cessation* **12**, 63–70 (2017).

[275] Silva, E. H., Lawler, S. & Langbecker, D. The effectiveness of mhealth for self-management in improving pain, psychological distress, fatigue, and sleep in cancer survivors: a systematic review. *Journal of Cancer Survivorship* **13**, 97–107 (2019).

[276] Palesh, O. *et al.* Secondary outcomes of a behavioral sleep intervention: A randomized clinical trial. *Health Psychology* **38**, 196 (2019).

[277] Mussa, B. M., Schauman, M., Kumar, V., Skaria, S. & Abusnana, S. Personalized intervention to improve stress and sleep patterns for glycemic control and weight management in obese emirati patients with type 2 diabetes: a randomized controlled clinical trial. *Diabetes, metabolic syndrome and obesity: targets and therapy* **12**, 991 (2019).

[278] Khosla, S. *et al.* Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement. *Journal of Clinical Sleep Medicine* **14**, 877–880 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29734997http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5940440.

[279] Tuominen, J., Peltola, K., Saaresranta, T. & Valli, K. Sleep parameter assessment accuracy of a consumer home sleep monitoring ballistocardiograph beddit sleep tracker: A validation study. *Journal of Clinical Sleep Medicine* **15**, 483–487 (2019).

[280] de Korte, E. M., Wiezer, N., Janssen, J. H., Vink, P. & Kraaij, W. Evaluating an mhealth app for health and well-being at work: mixed-method qualitative study. *JMIR mHealth and uHealth* **6**, e72 (2018).

[281] Sjövall, S. *et al. Coping with stress: Firstbeat Lifestyle Assessments for family workers*. Ph.D. thesis, Satakunta University of Applied Sciences, Satakunnan ammattikorkeakoulu (2015).

[282] Munzner, T. *Visualization Analysis and Design*. A.K. Peters visualization series (A K Peters, 2014). URL http://www.cs.ubc.ca/%7Etmm/vadbook/.

[283] YK, C. *et al.* Smartphone applications to support sleep self-management: review and evaluation. *J. Clin Sleep Med.* **14**, 1783,1790 (2018).

[284] Nonato, L. G. & Aupetit, M. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Trans. Vis. Comput. Graph.* **25**, 2650–2673 (2019).

[285] Fuster-Garcia, E., Bresó, A., Miranda, J. M. & García-Gómez, J. M. *Actigraphy Pattern Analysis for Outpatient Monitoring*, 3–17 (Springer New York, New York, NY, 2015). URL https://doi.org/10.1007/978-1-4939-1985-7_1.

[286] Liang, Z. *et al.* Sleepexplorer: a visualization tool to make sense of correlations between personal sleep data and contextual factors. *Personal and Ubiquitous Computing* **20** (2016).

[287] Duncan, M. *et al.* Activity trackers implement different behavior change techniques for activity, sleep, and sedentary behaviors. *Interact J Med Res* **6**, e13 (2017).

[288] Ravichandran, R., Sien, S.-W., Patel, S. N., Kientz, J. A. & Pina, L. R. Making sense of sleep sensors: How sleep sensing technologies support and undermine sleep health. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 6864–6875 (ACM, New York, NY, USA, 2017). URL http://doi.acm.org/10.1145/3025453.3025557.

[289] IEEE. *IEEE VIS 2015 Workshop on Personal Visualization: Exploring Data in Everyday Life*. URL http://www.vis4me.com/personalvis15/papers.html.

[290] Ryokai, K., Michahelles, F., Kritzler, M. & Syed, S. Communicating and interpreting wearable sensor data with health coaches. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, 221–224 (2015).

[291] Khairat, S. S. *et al.* The impact of visualization dashboards on quality of care and clinician satisfaction: Integrative literature review. *JMIR Hum Factors* **5**, e22 (2018). URL https://doi.org/10.2196/humanfactors.9328.

[292] Gewin, V. Data sharing: An open mind on open data. *Nature* **529**, 117–119 (2016).

[293] Dinov, I. D. Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. *GigaScience* **5**, 12 (2016).

[294] Turakhia, M. P. *et al.* Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The apple heart study. *American heart journal* **207**, 66–75 (2019).

[295] Dean, D. A. *et al.* Scaling up scientific discovery in sleep medicine: the national sleep research resource. *Sleep* **39**, 1151–1164 (2016).

[296] Lichstein, K. L. *et al.* Telehealth cognitive behavior therapy for co-occurring insomnia and depression symptoms in older adults. *Journal of clinical psychology* **69**, 1056–1065 (2013).

[297] Holmqvist, M., Vincent, N. & Walsh, K. Web-vs telehealth-based delivery of cognitive behavioral therapy for insomnia: a randomized controlled trial. *Sleep medicine* **15**, 187–195 (2014).

[298] van Drongelen, A. *et al.* Evaluation of an mhealth intervention aiming to improve health-related behavior and sleep and reduce fatigue among airline pilots. *Scandinavian journal of work, environment & health* 557–568 (2014).

[299] Babson, K. A., Ramo, D. E., Baldini, L., Vandrey, R. & Bonn-Miller, M. O. Mobile app-delivered cognitive behavioral therapy for insomnia: feasibility and initial efficacy among veterans with cannabis use disorders. *JMIR research protocols* **4** (2015).

[300] Shin, J. C., Kim, J. & Grigsby-Toussaint, D. Mobile phone interventions for sleep disorders and sleep quality: systematic review. *JMIR mHealth and uHealth* **5** (2017).

[301] Sáez, C. & García-Gómez, J. M. Kinematics of big biomedical data to characterize temporal variability and seasonality of data repositories: Functional data analysis of data temporal evolution over non-parametric statistical manifolds. *International Journal of Medical Informatics* **119**, 109–124 (2018). URL http://www.sciencedirect.com/science/article/pii/S138650561830563X.

[302] Sáez, C., Robles, M. & García-Gómez, J. M. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Statistical methods in medical research* **26**, 312–336 (2017).

[303] Mathews, S. C. *et al.* Digital health: a path to validation. *npj Digital Medicine* **2**, 38 (2019). URL http://www.nature.com/articles/s41746-019-0111-3.

[304] Grigsby-Toussaint, D. S. *et al.* Sleep apps and behavioral constructs: a content analysis. *Preventive medicine reports* **6**, 126–129 (2017).

[305] Fino, E. & Mazzetti, M. Monitoring healthy and disturbed sleep through smartphone applications: a review of experimental evidence. *Sleep and Breathing* 1–12 (2018).

[306] Piwek, L., Ellis, D. A., Andrews, S. & Joinson, A. The rise of consumer health wearables: promises and barriers. *PLoS Medicine* **13**, e1001953 (2016).

[307] Lauritzen, J., Munoz, A., Luis, J. S. & Civit, A. The usefulness of activity trackers in elderly with reduced mobility: a case study. *Studies in health technology and informatics* **192**, 759–762 (2013).

[308] Wilbanks, J. T. & Topol, E. J. Stop the privatization of health data. *Nature News* **535**, 345 (2016).

[309] Pfiffner, P. B., Pinyol, I., Natter, M. D. & Mandl, K. D. C3-pro: connecting researchkit to the health system using i2b2 and fhir. *PloS one* **11**, e0152722 (2016).

[310] Coravos, A., Khozin, S. & Mandl, K. D. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *npj Digital Medicine* **2**, 14 (2019). URL http://www.nature.com/articles/s41746-019-0090-4.

[311] Kay, M. *et al.* Lullaby: a capture & access system for understanding the sleep environment. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, 226–234 (ACM, 2012).

[312] Phillips, A. J. *et al.* Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing. *Scientific reports* **7**, 3216 (2017).

[313] Penzel, T., Fietze, I. & Veauthier, C. The need for a reliable sleep eeg biomarker. *Journal of Clinical Sleep Medicine* **13**, 771–772 (2017).

[314] Levendowski, D. J. *et al.* The accuracy, night-to-night variability, and stability of frontopolar sleep electroencephalography biomarkers. *Journal of Clinical Sleep Medicine* **13**, 791–803 (2017).

[315] Forner-Cordero, A., Umemura, G. S., Furtado, F. & Gonçalves, B. d. S. B. Comparison of sleep quality assessed by actigraphy and questionnaires to healthy subjects. *Sleep Science* **11**, 141 (2018).

[316] Schwartz, J. R. L. & Roth, T. Neurophysiology of sleep and wakefulness: basic science and clinical implications. *Current neuropharmacology* **6**, 367–78 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/19587857http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2701283.

[317] Imeri, L. & Opp, M. R. How (and why) the immune system makes us sleep (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19209176http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2839418.

[318] Ben Simon, E., Rossi, A., Harvey, A. G. & Walker, M. P. Overanxious and underslept. *Nature Human Behaviour* 1–11 (2019). URL http://www.nature.com/articles/s41562-019-0754-8.

[319] Fultz, N. E. *et al.* Coupled electrophysiological, hemodynamic, and cerebrospinal fluid oscillations in human sleep. *Science (New York, N.Y.)* **366**, 628–631 (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/31672896.

[320] Abdullah, S., Matthews, M., Murnane, E. L., Gay, G. & Choudhury, T. Towards circadian computing: "Early to bed and early to rise" makes some of us unhealthy and sleep deprived. In *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 673–684 (Association for Computing Machinery, Inc, 2014).

[321] Berry, R. B. *et al.* American Academy of Sleep Medicine. The AASM Manual for the Scoring of Sleep and Associated Events : Rules, Terminology, and Technical Specifications, Version 2.2. *American Academy of Sleep* **28**, 391–397 (2016). URL www.aasmnet.org.

[322] Girschik, J., Fritschi, L., Heyworth, J. & Waters, F. Validation of self-reported sleep against actigraphy. *Journal of epidemiology* **22**, 462–8 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22850546http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3798642.

[323] Sadeh, A., Hauri, P. J., Kripke, D. F. & Lavie, P. The role of actigraphy in the evaluation of sleep disorders. *Sleep* **18**, 288–302 (1995). URL https://academic.oup.com/sleep/article-abstract/18/4/288/2749735http://www.ncbi.nlm.nih.gov/pubmed/7618029.

[324] Sazonov, E., Sazonova, N., Schuckers, S., Neuman, M. & CHIME Study Group. Activity-based sleep-wake identification in infants. *Physiological measurement* **25**, 1291–304 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15535193.

[325] Lee, J.-M., Byun, W., Keill, A., Dinkel, D. & Seo, Y. Comparison of Wearable Trackers' Ability to Estimate Sleep. *International journal of environmental research and public health* **15**, 1265 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29914050http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6025478.

[326] Aktaruzzaman, M. *et al.* The addition of entropy-based regularity parameters improves sleep stage classification based on heart rate variability. *Medical and Biological Engineering and Computing* **53**, 415–425 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25690323http://link.springer.com/10.1007/s11517-015-1249-z.

[327] Bonnet, M. H. & Arand, D. L. Heart rate variability: sleep stage, time of night, and arousal influences. *Electroencephalography and Clinical Neurophysiology* **102**, 390–396 (1997). URL https://www.sciencedirect.com/science/article/pii/S0921884X96960701.

[328] Vanoli, E. *et al.* Heart Rate Variability During Specific Sleep Stages. *Circulation* **91**, 1918–1922 (1995). URL https://www.ahajournals.org/doi/10.1161/01.CIR.91.7.1918.

[329] Elsenbruch, S., Harnish, M. J. & Orr, W. C. Heart Rate Variability During Waking and Sleep in Healthy Males and Females. *Sleep* **22**, 1067–1071 (1999). URL http://www.ncbi.nlm.nih.gov/pubmed/10617167https://academic.oup.com/sleep/article-lookup/doi/10.1093/sleep/22.8.1067.

[330] Daskalova, N., Lee, B., Huang, J., Ni, C. & Lundin, J. Investigating the Effectiveness of Cohort-Based Sleep Recommendations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **2**, 1–19 (2018).

[331] Yuda, E. *et al.* Sleep Stage Classification by a Combination of Actigraphic and Heart Rate Signals. *Journal of Low Power Electronics and Applications* **7**, 28 (2017). URL http://www.mdpi.com/2079-9268/7/4/28.

[332] Hees, V. T. v. *et al.* Estimating sleep parameters using an accelerometer without sleep diary. *Scientific Reports* **8**, 12975 (2018). URL https://www.biorxiv.org/content/early/2018/02/01/257972.

[333] Dean, D. A. *et al.* Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. *Sleep* **39**, 1151–1164 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27070134http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4835314https://academic.oup.com/sleep/article-lookup/doi/10.5665/sleep.5774.

[334] Zhang, G.-Q. *et al.* The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association* **25**, 1351–1358 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29860441http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6188513https://academic.oup.com/jamia/article/25/10/1351/5026200.

[335] Tobaldini, E. *et al.* Heart rate variability in normal and pathological sleep. *Frontiers in Physiology* **4**, 1–11 (2013). URL http://journal.frontiersin.org/article/10.3389/fphys.2013.00294/abstract.

[336] Snyder, F., Hobson, J. A., Morrison, D. F. & Goldfrank, F. Changes in respiration, heart rate, and systolic blood pressure in human sleep. *Journal of Applied Physiology* **19**, 417–422 (1964). URL http://www.ncbi.nlm.nih.gov/pubmed/14174589https://www.physiology.org/doi/10.1152/jappl.1964.19.3.417.

[337] Kirby, D. A. & Verrier, R. L. Differential effects of sleep stage on coronary hemodynamic function. *American Journal of Physiology-Heart and Circulatory Physiology* **256**, H1378–H1383 (1989). URL http://www.ncbi.nlm.nih.gov/pubmed/2719135https://www.physiology.org/doi/10.1152/ajpheart.1989.256.5.H1378.

[338] Hornyak, M., Cejnar, M., Elam, M., Matousek, M. & Wallin, B. G. Sympathetic muscle nerve activity during sleep in man. *Brain* **114 (Pt 3**, 1281–95 (1991). URL http://www.ncbi.nlm.nih.gov/pubmed/2065250https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/114.3.1281.

[339] Somers, V. K., Dyken, M. E., Mark, A. L. & Abboud, F. M. Sympathetic-Nerve Activity during Sleep in Normal Subjects. *New England Journal of Medicine* **328**, 303–307 (1993). URL http://www.ncbi.nlm.nih.gov/pubmed/8419815http://www.nejm.org/doi/abs/10.1056/NEJM199302043280502.

[340] Boudreau, P., Yeh, W.-H., Dumont, G. A. & Boivin, D. B. Circadian Variation of Heart Rate Variability Across Sleep Stages. *Sleep* **36**, 1919–1928 (2013).

[341] Radha, M. *et al.* Sleep stage classification from heart-rate variability using long short-term memory neural networks. *Scientific Reports* **9**, 1–11 (2019).

[342] Singh, U., Chauhan, S., Krishnamachari, A. & Vig, L. Ensemble of deep long short term memory networks for labelling origin of replication sequences. In *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*, 1–7 (IEEE, 2015). URL http://ieeexplore.ieee.org/document/7344871/.

[343] Guan, Y. & Plötz, T. Ensembles of Deep LSTM Learners for Activity Recognition using Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1**, 1–28 (2017).

[344] Koley, B. & Dey, D. An ensemble system for automatic sleep stage classification using single channel EEG signal. *Computers in Biology and Medicine* **42**, 1186–1195 (2012).

[345] Alickovic, E. & Subasi, A. Ensemble SVM Method for Automatic Sleep Stage Classification. *IEEE Transactions on Instrumentation and Measurement* **67**, 1258–1265 (2018). URL https://ieeexplore.ieee.org/document/8292946/.

[346] Stephansen, J. B. *et al.* Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications* **9**, 5229 (2018). URL http://www.nature.com/articles/s41467-018-07229-3.

[347] Melgani, F. & Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing* **42**, 1778–1790 (2004).

[348] Phan, H., Andreotti, F., Cooray, N., Chen, O. Y. & De Vos, M. Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification. *IEEE Transactions on Biomedical Engineering* **66**, 1285–1296 (2019).

[349] Bild, D. E. *et al.* Multi-Ethnic Study of Atherosclerosis: Objectives and design. *American Journal of Epidemiology* **156**, 871–881 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/12397006https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwf113.

[350] Varri, A., Kemp, B., Penzel, T. & Schlogl, A. Standards for biomedical signal databases. *IEEE Engineering in Medicine and Biology Magazine* **20**, 33–37 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11446207http://ieeexplore.ieee.org/document/932722/.

[351] Tanaka, H., Monahan, K. D. & Seals, D. R. Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology* **37**, 153–156 (2001).

[352] Malik, M. Heart Rate Variability. *Annals of Noninvasive Electrocardiology* **1**, 151–181 (1996). URL http://doi.wiley.com/10.1111/j.1542-474X.1996.tb00275.x.

[353] Tilmanne, J., Urbain, J., Kothare, M. V., Wouwer, A. V. & Kothare, S. V. Algorithms for sleep-wake identification using actigraphy: A comparative study and new results. *Journal of Sleep Research* **18**, 85–98 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19250177.

[354] Kripke, D. F. *et al.* Wrist actigraphic scoring for sleep laboratory patients: Algorithm development. *Journal of Sleep Research* **19**, 612–619 (2010).

[355] Ponnusamy, A., Marques, J. L. B. & Reuber, M. Comparison of heart rate variability parameters during complex partial seizures and psychogenic nonepileptic seizures. *Epilepsia* **53**, 1314–21 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22642646.

[356] Patel, S. R. *et al.* Reproducibility of a Standardized Actigraphy Scoring Algorithm for Sleep in a US Hispanic/Latino Population. *Sleep* **38**, 1497–1503 (2015).

[357] Colten, H. R. & Altevogt, B. M. *Sleep disorders and sleep deprivation: An unmet public health problem* (National Academies Press, 2006).

[358] Bagaveyev, S. & Cook, D. J. Designing and evaluating active learning methods for activity recognition. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 469–478 (2014).

[359] Guo, H., Chen, L., Peng, L. & Chen, G. Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble. In *UbiComp 2016*, 1112–1123 (Association for Computing Machinery, Inc, 2016).

[360] Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv preprint* 1–9 (2018). URL http://github.com/slundberg/shaphttp://arxiv.org/abs/1802.03888.

[361] Saad, M. *et al.* Using heart rate profiles during sleep as a biomarker of depression. *BMC Psychiatry* **19**, 1–11 (2019). URL https://github.com/slundberg/shap.

[362] Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *ArXiv* **abs/1802.03888**, 1–9 (2018).

[363] Thomson, E. A. *et al.* Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities. *Journal of Sports Sciences* **37**, 1411–1419 (2019). URL https://www.tandfonline.com/doi/full/10.1080/02640414.2018.1560644.

[364] Simon, E. B., Rossi, A., Harvey, A. G. & Walker, M. P. Overanxious and underslept. *Nature Human Behaviour* **4**, 100–110 (2020).

[365] Vaswani, A. *et al.* Attention is All you Need. In *Advances in neural information processing systems*, 5998–6008 (Curran Associates, Long Beach, CA, United States, 2017). URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[366] Hobson, J. A., McCarley, R. W. & Wyzinski, P. W. Sleep cycle oscillation: Reciprocal discharge by two brainstem neuronal groups. *Science* **189**, 55–58 (1975). URL http://www.jstor.org/stable/1740806.

[367] Baroni, A., Bruzzese, J. M., Di Bartolo, C. A. & Shatkin, J. P. Fitbit Flex: an unreliable device for longitudinal sleep measures in a non-clinical population (2016).

[368] Rasch, B. & Born, J. About sleep's role in memory. *Physiological reviews* **93**, 681–766 (2013).

[369] Schwartz, J. R. & Roth, T. Neurophysiology of sleep and wakefulness: basic science and clinical implications. *Current neuropharmacology* **6**, 367–378 (2008).

[370] Imeri, L. & Opp, M. R. How (and why) the immune system makes us sleep. *Nature Reviews Neuroscience* **10**, 199–210 (2009).

[371] Xie, L. *et al.* Sleep drives metabolite clearance from the adult brain. *science* **342**, 373–377 (2013).

[372] Adam, K. & Oswald, I. Sleep helps healing. *British medical journal (Clinical research ed.)* **289**, 1400 (1984).

[373] Benington, J. H. & Heller, H. C. Restoration of brain energy metabolism as the function of sleep. *Progress in neurobiology* **45**, 347–360 (1995).

[374] Bertisch, S. M. *et al.* Insomnia with objective short sleep duration and risk of incident cardiovascular disease and all-cause mortality: Sleep heart health study. *Sleep* **41**, zsy047 (2018).

[375] Dawson, D. & Reid, K. Fatigue, alcohol and performance impairment. *Nature* **388**, 235–235 (1997).

[376] Van Cauter, E., Spiegel, K., Tasali, E. & Leproult, R. Metabolic consequences of sleep and sleep loss. *Sleep medicine* **9**, S23–S28 (2008).

[377] St-Onge, M.-P. *et al.* Sleep duration and quality: impact on lifestyle behaviors and cardiometabolic health: a scientific statement from the american heart association. *Circulation* **134**, e367–e386 (2016).

[378] Agnew Jr, H., Webb, W. B. & Williams, R. L. The first night effect: an eeg studyof sleep. *Psychophysiology* **2**, 263–266 (1966).

[379] Ancoli-Israel, S. *et al.* The role of actigraphy in the study of sleep and circadian rhythms. *Sleep* **26**, 342–392 (2003).

[380] Marino, M. *et al.* Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep* **36**, 1747–1755 (2013).

[381] Kupfer, D. J., Detre, T. P., Foster, G., Tucker, G. J. & Delgado, J. The application of delgado's telemetric mobility recorder for human studies. *Behavioral biology* **7**, 585–590 (1972).

[382] Sadeh, A. & Acebo, C. The role of actigraphy in sleep medicine. *Sleep medicine reviews* **6**, 113–124 (2002).

[383] de Souza, L. *et al.* Further validation of actigraphy for sleep studies. *Sleep* **26**, 81–85 (2003).

[384] Sazonov, E. *et al.* Activity-based sleep–wake identification in infants. *Physiological measurement* **25**, 1291 (2004).

[385] Sadeh, A., Sharkey, M. & Carskadon, M. A. Activity-based sleep-wake identification: an empirical test of methodological issues. *Sleep* **17**, 201–207 (1994).

[386] Tilmanne, J., Urbain, J., Kothare, M. V., Wouwer, A. V. & Kothare, S. V. Algorithms for sleep–wake identification using actigraphy: a comparative study and new results. *Journal of sleep research* **18**, 85–98 (2009).

[387] Kripke, D. F. *et al.* Wrist actigraphic scoring for sleep laboratory patients: algorithm development. *Journal of sleep research* **19**, 612–619 (2010).

[388] Blood, M. L., Sack, R. L., Percy, D. C. & Pen, J. C. A comparison of sleep detection by wrist actigraphy, behavioral response, and polysomnography. *Sleep* **20**, 388–395 (1997).

[389] Paquet, J., Kawinska, A. & Carrier, J. Wake detection capacity of actigraphy during sleep. *Sleep* **30**, 1362–1369 (2007).

[390] van Hees, V. T. *et al.* Estimating sleep parameters using an accelerometer without sleep diary. *Scientific reports* **8**, 1–11 (2018).

[391] of Us Research Program Investigators, A. The "all of us" research program. *New England Journal of Medicine* **381**, 668–676 (2019).

[392] Walch, O., Huang, Y., Forger, D. & Goldstein, C. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep* **42**, zsz180 (2019).

[393] Roberts, D. M., Schade, M. M., Mathew, G. M., Gartenberg, D. & Buxton, O. M. Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography. *Sleep* (2020).

[394] de Zambotti, M., Trinder, J., Silvani, A., Colrain, I. M. & Baker, F. C. Dynamic coupling between the central and autonomic nervous systems during sleep: a review. *Neuroscience & Biobehavioral Reviews* **90**, 84–103 (2018).

[395] Lauderdale, D. S., Knutson, K. L., Yan, L. L., Liu, K. & Rathouz, P. J. Self-reported and measured sleep duration: how similar are they? *Epidemiology* 838–845 (2008).

[396] O'Connor, L., Brage, S., Griffin, S. J., Wareham, N. J. & Forouhi, N. G. The cross-sectional association between snacking behaviour and measures of adiposity: the fenland study, uk. *British journal of nutrition* **114**, 1286–1293 (2015).

[397] White, T. *et al.* Estimating energy expenditure from wrist and thigh accelerometry in free-living adults: a doubly labelled water study. *International Journal of Obesity* **43**, 2333–2342 (2019).

[398] Chen, X. *et al.* Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep* **38**, 877–888 (2015).

[399] Goldberger, A. L. *et al.* Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation* **101**, e215–e220 (2000).

[400] Rossi, A. *et al.* Multilevel monitoring of activity and sleep in healthy people (version 1.0.0). *PhysioNet* (2020).

[401] Brage, S., Brage, N., Franks, P. W., Ekelund, U. & Wareham, N. J. Reliability and validity of the combined heart rate and movement sensor actiheart. *European journal of clinical nutrition* **59**, 561–570 (2005).

[402] Stegle, O., Fallert, S. V., MacKay, D. J. & Brage, S. Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering* **55**, 2143–2151 (2008).

[403] Van Hees, V. T. *et al.* Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents. *Journal of Applied Physiology* **117**, 738–744 (2014).

[404] Lukowicz, P., Junker, H. & Tröster, G. Automatic calibration of body worn acceleration sensors. In *International Conference on Pervasive Computing*, 176–181 (Springer, 2004).

[405] Varri, A., Kemp, B., Penzel, T. & Schlogl, A. Standards for biomedical signal databases. *IEEE Engineering in Medicine and Biology Magazine* **20**, 33–37 (2001).

[406] Jackson, C. L., Patel, S. R., Jackson, W. B., Lutsey, P. L. & Redline, S. Agreement between self-reported and objectively measured sleep duration among white, black, hispanic, and chinese adults in the united states: Multi-ethnic study of atherosclerosis. *Sleep* **41**, zsy057 (2018).

[407] Karjalainen, J. & Viitasalo, M. Fever and cardiac rhythm. *Archives of internal medicine* **146**, 1169–1171 (1986).

[408] Ryan, J. & Howes, L. Relations between alcohol consumption, heart rate, and heart rate variability in men. *Heart* **88**, 641–642 (2002).

[409] Vrijkotte, T. G., Van Doornen, L. J. & De Geus, E. J. Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability. *Hypertension* **35**, 880–886 (2000).

[410] Radin, J. M., Wineinger, N. E., Topol, E. J. & Steinhubl, S. R. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the usa: a population-based study. *The Lancet Digital Health* (2020).

[411] Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20**, 37–46 (1960).

[412] Freund, J. E. *Modern Elementary Statistics* (Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988).

[413] Snyder, F., Hobson, J. A., Morrison, D. F. & Goldfrank, F. Changes in respiration, heart rate, and systolic blood pressure in human sleep. *Journal of applied physiology* **19**, 417–422 (1964).

[414] Bliwise, D. L. Invited commentary: cross-cultural influences on sleep—broadening the environmental landscape. *American journal of epidemiology* **168**, 1365–1366 (2008).

[415] Wesselius, H. M. *et al.* Quality and Quantity of Sleep and Factors Associated With Sleep Disturbance in Hospitalized Patients. *JAMA Internal Medicine* **178**, 1201–1208 (2018). URL https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2687528. Publisher: American Medical Association.

[416] Dishman, R. K. *et al.* Decline in Cardiorespiratory Fitness and Odds of Incident Sleep Complaints. *Medicine & Science in Sports & Exercise* **47**, 960–966 (2015). URL https://journals.lww.com/acsm-msse/Fulltext/2015/05000/Decline_in_Cardiorespiratory_Fitness_and_Odds_of.10.aspx.

[417] Depner, C. M. *et al.* Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *sleep* **43**, zsz254 (2020).

[418] da Silva, I. C. *et al.* Physical activity levels in three brazilian birth cohorts as assessed with raw triaxial wrist accelerometry. *International journal of epidemiology* **43**, 1959–1968 (2014).

[419] Menai, M. *et al.* Accelerometer assessed moderate-to-vigorous physical activity and successful ageing: results from the whitehall ii study. *Scientific reports* **7**, 45772 (2017).

[420] Chuah, S. H.-W. *et al.* Wearable technologies: The role of usefulness and visibility in smartwatch adoption. *Computers in Human Behavior* **65**, 276–284 (2016).

[421] Lamkin, P. Wearable tech market to be worth 34 billion by 2020. *Forbes* 17 (2016).

[422] Cho, M. & Kim, J.-Y. Changes in physical fitness and body composition according to the physical activities of korean adolescents. *Journal of exercise rehabilitation* **13**, 568 (2017).

[423] Steeves, J. A. *et al.* Ability of thigh-worn actigraph and activpal monitors to classify posture and motion. *Medicine and science in sports and exercise* **47**, 952 (2015).

[424] Rowlands, A. V. *et al.* Sedentary sphere: wrist-worn accelerometer-brand independent posture classification. *Medicine & Science in Sports & Exercise* **48**, 748–754 (2016).

[425] Tremblay, M. S. *et al.* Sedentary behavior research network (sbrn)–terminology consensus project process and outcome. *International Journal of Behavioral Nutrition and Physical Activity* **14**, 75 (2017).

[426] Matthews, C. E. *et al.* Amount of time spent in sedentary behaviors in the united states, 2003–2004. *American journal of epidemiology* **167**, 875–881 (2008).

[427] Ekelund, U. *et al.* Does physical activity attenuate, or even eliminate, the detrimental association of sitting time with mortality? a harmonised meta-analysis of data from more than 1 million men and women. *The Lancet* **388**, 1302–1310 (2016).

[428] Patterson, R. *et al.* Sedentary behaviour and risk of all-cause, cardiovascular and cancer mortality, and incident type 2 diabetes: a systematic review and dose response meta-analysis (2018).

[429] van der Ploeg, H. P. & Hillsdon, M. Is sedentary behaviour just physical inactivity by another name? *International Journal of Behavioral Nutrition and Physical Activity* **14**, 142 (2017).

[430] Brage, S. *et al.* Hierarchy of individual calibration levels for heart rate and accelerometry to measure physical activity. *Journal of Applied Physiology* **103**, 682–692 (2007).

[431] Strath, S. J., Brage, S. & Ekelund, U. Integration of physiological and accelerometer data to improve physical activity assessment. *Medicine & Science in Sports & Exercise* **37**, S563–S571 (2005).

[432] Thompson, D., Batterham, A. M., Bock, S., Robson, C. & Stokes, K. Assessment of low-to-moderate intensity physical activity thermogenesis in young adults using synchronized heart rate and accelerometry with branched-equation modeling. *The Journal of nutrition* **136**, 1037–1042 (2006).

[433] Villars, C. *et al.* Validity of combining heart rate and uniaxial acceleration to measure free-living physical activity energy expenditure in young men. *Journal of applied physiology* **113**, 1763–1771 (2012).

[434] Brage, S. *et al.* Estimation of free-living energy expenditure by heart rate and movement sensing: a doubly-labelled water study. *PloS one* **10** (2015).

[435] van Hees, V. T. *et al.* Estimation of daily energy expenditure in pregnant and non-pregnant women using a wrist-worn tri-axial accelerometer. *PloS one* **6** (2011).

[436] Arora, T. *et al.* The complexity of obesity in uk adolescents: relationships with quantity and type of technology, sleep duration and quality, academic performance and aspiration. *Pediatric Obesity* **8**, 358–366 (2013).

[437] Brage, S. *et al.* Descriptive epidemiology of energy expenditure in the uk: findings from the national diet and nutrition survey 2008–15. *International Journal of Epidemiology* (2020).

[438] O'Donnell, J. *et al.* Automated detection of sleep-boundary times using wrist-worn accelerometry. *BioRxiv* 225516 (2017).

[439] Blalock, D. W. & Guttag, J. V. Extract: Strong examples from weakly-labeled sensor data. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 799–804 (IEEE, 2016).

[440] Jones, A. M. & Carter, H. The effect of endurance training on parameters of aerobic fitness. *Sports medicine* **29**, 373–386 (2000).

[441] Ellestad, M. H. & Wan, M. Predictive implications of stress testing. follow-up of 2700 subjects after maximum treadmill stress testing. *Circulation* **51**, 363–369 (1975).

[442] Savonen, K. P. *et al.* Heart rate response during exercise test and cardiovascular mortality in middle-aged men. *European heart journal* **27**, 582–588 (2006).

[443] Naeinia, E. K., Azimib, I., Rahmania, A. M., Liljebergb, P. & Dutta, N. A real-time ppg quality assessment approach for healthcare internet-of-things. *Procedia Computer Science* **151**, 551–558 (2019).

[444] Ang, W. T., Khosla, P. K. & Riviere, C. N. Nonlinear regression model of a low-g mems accelerometer. *IEEE Sensors Journal* **7**, 81–88 (2006).

[445] Bulling, A., Blanke, U. & Schiele, B. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)* **46**, 1–33 (2014).

[446] Krishnan, A., Sharma, A. & Sundaram, H. Insights from the long-tail: Learning latent representations of online user behavior in the presence of skew and sparsity. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 297–306 (2018).

[447] Cho, K. *et al.* Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP* (2014).

[448] Bai, S., Kolter, J. Z. & Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[449] Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* **5**, 157–166 (1994).

[450] Hammerla, N. Y., Halloran, S. & Plötz, T. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 1533–1540 (AAAI Press, 2016).

[451] Yang, J., Nguyen, M. N., San, P. P., Li, X. L. & Krishnaswamy, S. Deep convolutional neural networks on multichannel time series for human activity recognition. In *IJCAI* (2015).

[452] Ma, H., Li, W., Zhang, X., Gao, S. & Lu, S. Attnsense: multi-level attention mechanism for multimodal human activity recognition. In *IJCAI* (2019).

[453] Shin, H.-C. *et al.* Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging* **35**, 1285–1298 (2016).

[454] Chen, H., Lundberg, S., Erion, G., Kim, J. H. & Lee, S.-I. Deep transfer learning for physiological signals. *arXiv preprint arXiv:2002.04770* (2020).

[455] Lan, Z. *et al.* Albert: A lite bert for self-supervised learning of language representations. *ICLR* (2020).

[456] Owens, A., Wu, J., McDermott, J. H., Freeman, W. T. & Torralba, A. Ambient sound provides supervision for visual learning. In *ECCV* (2016).

[457] Althoff, T. *et al.* Large-scale physical activity data reveal worldwide activity inequality. *Nature* **547**, 336 (2017).

[458] Mandsager, K. *et al.* Association of cardiorespiratory fitness with long-term mortality among adults undergoing exercise treadmill testing. *JAMA network open* **1**, e183605–e183605 (2018).

[459] Ni, J., Muhlstein, L. & McAuley, J. Modeling heart rate and activity data for personalized fitness recommendation. In *WWW* (2019).

[460] McConville, R. *et al.* Online heart rate prediction using acceleration from a wrist worn wearable. *arXiv preprint arXiv:1807.04667* (2018).

[461] Jenni, S. & Favaro, P. Self-supervised feature learning by learning to spot artifacts. In *CVPR* (2018).

[462] Sarkar, P. & Etemad, A. Self-supervised learning for ecg-based emotion recognition. *arXiv preprint arXiv:1910.07497* (2019).

[463] Hallgrímsson, H. T., Jankovic, F., Althoff, T. & Foschini, L. Learning individualized cardiovascular responses from large-scale wearable sensors data. *NIPS ML4H workshop* (2018).

[464] Wu, X., Huang, C., Roblesgranda, P. & Chawla, N. Representation learning on variable length and incomplete wearable-sensory time series. *arXiv preprint arXiv:2002.03595* (2020).

[465] Sanchez-Lengeling, B. *et al.* Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:1910.10685* (2019).

[466] Le, Q. & Mikolov, T. Distributed representations of sentences and documents. In *ICML* (2014).

[467] Chen, R. *et al.* Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *KDD* (2019).

[468] Yao, S., Hu, S., Zhao, Y., Zhang, A. & Abdelzaher, T. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th International Conference on World Wide Web*, 351–360 (International World Wide Web Conferences Steering Committee, 2017).

[469] Tanaka, H., Monahan, K. D. & Seals, D. R. Age-predicted maximal heart rate revisited. *Journal of the american college of cardiology* **37**, 153–156 (2001).

[470] Rodrigues, F. & Pereira, F. C. Beyond expectation: Deep joint mean and quantile regression for spatio-temporal problems. *arXiv preprint arXiv:1808.08798* (2018).

[471] Dabney, W., Rowland, M., Bellemare, M. G. & Munos, R. Distributional reinforcement learning with quantile regression. In *AAAI* (2018).

[472] Brage, S. *et al.* Branched equation modeling of simultaneous accelerometry and heart rate monitoring improves estimate of directly measured physical activity energy expenditure. *Journal of applied physiology* **96**, 343–351 (2004).

[473] Chakraborty, D. & Elzarka, H. Advanced machine learning techniques for building performance simulation: a comparative analysis. *Journal of Building Performance Simulation* **12**, 193–207 (2019).

[474] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[475] Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *KDD* (2016).

[476] Fox, K. *et al.* Resting heart rate in cardiovascular disease. *Journal of the American College of Cardiology* **50**, 823–830 (2007).

[477] Jaques, N., Taylor, S., Sano, A., Picard, R. *et al.* Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *IJCAI 2017 Workshop on artificial intelligence in affective computing*, 17–33 (2017).

[478] Manini, T. M. *et al.* Daily activity energy expenditure and mortality among older adults. *Jama* **296**, 171–179 (2006).

[479] Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[480] Maaten, L. v. d. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9**, 2579–2605 (2008).

[481] Lynch, J. *et al.* Moderately intense physical activities and high levels of cardiorespiratory fitness reduce the risk of non-insulin-dependent diabetes mellitus in middle-aged men. *Archives of internal medicine* **156**, 1307–1314 (1996).

[482] Lakka, T. A. *et al.* Relation of leisure-time physical activity and cardiorespiratory fitness to the risk of acute myocardial infarction in men. *New England Journal of Medicine* **330**, 1549–1554 (1994).

[483] Myers, J. *et al.* Exercise capacity and mortality among men referred for exercise testing. *New England journal of medicine* **346**, 793–801 (2002).

[484] Ekelund, L.-G. *et al.* Physical fitness as a predictor of cardiovascular mortality in asymptomatic north american men. *New England Journal of Medicine* **319**, 1379–1384 (1988).

[485] Schuch, F. B. *et al.* Are lower levels of cardiorespiratory fitness associated with incident depression? a systematic review of prospective cohort studies. *Preventive Medicine* **93**, 159–165 (2016).

[486] Blair, S. N. *et al.* Physical fitness and all-cause mortality: a prospective study of healthy men and women. *Jama* **262**, 2395–2401 (1989).

[487] Laukkanen, J. A., Kurl, S., Salonen, R., Rauramaa, R. & Salonen, J. T. The predictive value of cardiorespiratory fitness for cardiovascular events in men with various risk profiles: a prospective population-based cohort study. *European heart journal* **25**, 1428–1437 (2004).

[488] Ross, R. *et al.* Importance of assessing cardiorespiratory fitness in clinical practice: a case for fitness as a clinical vital sign: a scientific statement from the american heart association. *Circulation* **134**, e653–e699 (2016).

[489] Kokkinos, P. F., Faselis, C., Myers, J., Panagiotakos, D. & Doumas, M. Interactive effects of fitness and statin treatment on mortality risk in veterans with dyslipidaemia: a cohort study. *The Lancet* **381**, 394–399 (2013).

[490] Lloyd-Jones, D. M. *et al.* Defining and setting national goals for cardiovascular health promotion and disease reduction: the american heart association's strategic impact goal through 2020 and beyond. *Circulation* **121**, 586–613 (2010).

[491] Swain, D. P., Brawner, C. A., of Sports Medicine, A. C. *et al. ACSM's resource manual for guidelines for exercise testing and prescription* (Wolters Kluwer Health/Lippincott Williams & Wilkins, 2014).

[492] Davis, J. Direct determination of aerobic power. *Physiological assessment of human fitness* 9–17 (1995).

[493] Abut, F., Akay, M. F. & George, J. Developing new vo2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection. *Computers in biology and medicine* **79**, 182–192 (2016).

[494] Davies, C. Limitations to the prediction of maximum oxygen intake from cardiac frequency measurements. *Journal of Applied Physiology* **24**, 700–706 (1968).

[495] Noonan, V. & Dean, E. Submaximal exercise testing: clinical application and interpretation. *Physical therapy* **80**, 782–807 (2000).

[496] Turjanmaa, V., Kalli, S., Sydänmaa, M. & Uusitalo, A. Short-term variability of systolic blood pressure and heart rate in normotensive subjects. *Clinical Physiology* **10**, 389–401 (1990).

[497] Oja, P. Dose response between total volume of physical activity and health and fitness. *Medicine and science in sports and exercise* **33**, S428–37 (2001).

[498] LaMonte, M. J. & Blair, S. N. Physical activity, cardiorespiratory fitness, and adiposity: contributions to disease risk. *Current Opinion in Clinical Nutrition & Metabolic Care* **9**, 540–546 (2006).

[499] Nauman, J., Aspenes, S. T., Nilsen, T. I. L., Vatten, L. J. & Wisløff, U. A prospective population study of resting heart rate and peak oxygen uptake (the hunt study, norway). *PloS one* **7**, e45021 (2012).

[500] Cao, Z.-B. *et al.* Predicting v˙ o2max with an objectively measured physical activity in japanese women. *Medicine & Science in Sports & Exercise* **42**, 179–186 (2010).

[501] Jurca, R. *et al.* Assessing cardiorespiratory fitness without performing exercise testing. *American journal of preventive medicine* **29**, 185–193 (2005).

[502] Nes, B. M. *et al.* Estimating v˙ o2peak from a nonexercise prediction model: the hunt study, norway. *Medicine & Science in Sports & Exercise* **43**, 2024–2030 (2011).

[503] PLASQUI, G. & WESTERTERP, K. R. Accelerometry and heart rate as a measure of physical fitness: proof of concept. *Medicine & Science in Sports & Exercise* **37**, 872–876 (2005).

[504] Passler, S., Bohrer, J., Blöchinger, L. & Senner, V. Validity of wrist-worn activity trackers for estimating vo2max and energy expenditure. *International journal of environmental research and public health* **16**, 3037 (2019).

[505] Shcherbina, A. *et al.* Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of personalized medicine* **7**, 3 (2017).

[506] Boudreaux, B. D. *et al.* Validity of wearable activity monitors during cycling and resistance exercise. *Medicine and science in sports and exercise* **50**, 624–633 (2018).

[507] Esco, M. R., Mugu, E. M., Williford, H. N., McHugh, A. N. & Bloomquist, B. E. Cross-validation of the polar fitness test via the polar f11 heart rate monitor in predicting vo 2 max. *Journal of Exercise Physiology Online* **14** (2011).

[508] Cooper, K. D. & Shafer, A. B. Validity and reliability of the polar a300's fitness test feature to predict vo2max. *International journal of exercise science* **12**, 393 (2019).

[509] Lucio, N. D. *et al.* Accuracy of fitbit charge 2 at estimating vo2max, calories, and steps on a treadmill. In *International Journal of Exercise Science: Conference Proceedings*, vol. 2, 11 (2018).

[510] Altini, M., Penders, J. & Amft, O. Personalizing energy expenditure estimation using a cardiorespiratory fitness predicate. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, 65–72 (IEEE, 2013).

[511] RENNIE, K. L., HENNINGS, S. J., Mitchell, J. & WAREHAM, N. J. Estimating energy expenditure by heart-rate monitoring without individual calibration. *Medicine & Science in Sports & Exercise* **33**, 939–945 (2001).

[512] Uth, N., Sørensen, H., Overgaard, K. & Pedersen, P. K. Estimation of vo 2max from the ratio between hr max and hr rest–the heart rate ratio method. *European journal of applied physiology* **91**, 111–115 (2004).

[513] Cao, Z.-B. *et al.* Prediction of vo 2max with daily step counts for japanese adult women. *European journal of applied physiology* **105**, 289–296 (2009).

[514] Beltrame, T., Amelard, R., Wong, A. & Hughson, R. Prediction of oxygen uptake dynamics by machine learning analysis of wearable sensors during activities of daily living. *Scientific reports* **7**, 45738 (2017).

[515] Gonzales, T. I. *et al.* Estimating maximal oxygen consumption from heart rate response to submaximal ramped treadmill test. *medRxiv* (2020).

[516] Assah, F. K. *et al.* Accuracy and validity of a combined heart rate and motion sensor for the measurement of free-living physical activity energy expenditure in adults in cameroon. *International journal of epidemiology* **40**, 112–120 (2011).

[517] Christ, M., Braun, N., Neuffer, J. & Kempa-Liehr, A. W. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing* (2018).

[518] Mazess, R. B., Barden, H. S., Bisek, J. P. & Hanson, J. Dual-energy x-ray absorptiometry for total-body and regional bone-mineral and soft-tissue composition. *The American journal of clinical nutrition* **51**, 1106–1112 (1990).

[519] Dove, E. S. & Chen, J. Should consent for data processing be privileged in health research? a comparative legal analysis. *International Data Privacy Law* (2020).

[520] Jain, S. H., Powers, B. W., Hawkins, J. B. & Brownstein, J. S. The digital phenotype. *Nature biotechnology* **33**, 462–463 (2015).

[521] Onnela, J.-P. & Rauch, S. L. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* **41**, 1691–1696 (2016).

[522] White, R. W. & Horvitz, E. Population-scale hand tremor analysis via anonymized mouse cursor signals. *NPJ digital medicine* **2**, 1–7 (2019).

[523] Pasquale, F. Grand bargains for big data: The emerging law of health information. *Md. L. Rev.* **72**, 682 (2012).

[524] Troiano, A. Wearables and personal health data: putting a premium on your privacy. *Brook. L. Rev.* **82**, 1715 (2016).

[525] Wang, R. *et al.* Accuracy of wrist-worn heart rate monitors. *Jama cardiology* **2**, 104–106 (2017).

[526] Ajunwa, I. Algorithms at work: productivity monitoring applications and wearable technology as the new data-centric research agenda for employment and labor law. . *Louis ULJ* **63**, 21 (2018).

[527] Montgomery, K., Chester, J. & Kopp, K. Health wearables: ensuring fairness, preventing discrimination, and promoting equity in an emerging internet-of-things environment. *Journal of Information Policy* **8**, 34–77 (2018).

[528] Piwek, L., Ellis, D. A., Andrews, S. & Joinson, A. The rise of consumer health wearables: promises and barriers. *PLoS medicine* **13**, e1001953 (2016).

[529] Roberts, J. L. & Hawkins, J. When health tech companies change their terms of service. *Science* **367**, 745–746 (2020).

[530] Huckvale, K., Torous, J. & Larsen, M. E. Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. *JAMA network open* **2**, e192542–e192542 (2019).

[531] Viertler, C. & Zatloukal, K. Biobanking and biomolecular resources research infrastructure (bbmri). implications for pathology. *Der Pathologe* **29**, 210–213 (2008).

[532] Cline, M. S. *et al.* Brca challenge: Brca exchange as a global resource for variants in brca1 and brca2. *PLoS genetics* **14**, e1007752 (2018).

[533] Landrum, M. J. *et al.* Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic acids research* **46**, D1062–D1067 (2018).

[534] Karczewski, K. & Francioli, L. The genome aggregation database (gnomad). *MacArthur Lab* (2017).

[535] Lappalainen, I. *et al.* The european genome-phenome archive of human data consented for biomedical research. *Nature genetics* **47**, 692–695 (2015).

[536] Tryka, K. A. *et al.* Ncbi's database of genotypes and phenotypes: dbgap. *Nucleic acids research* **42**, D975–D979 (2014).

[537] Stenson, P. D. *et al.* The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human genetics* **133**, 1–9 (2014).

[538] Arias, J. J., Pham-Kanter, G. & Campbell, E. G. The growth and gaps of genetic data sharing policies in the united states. *Journal of Law and the Biosciences* **2**, 56–68 (2015).

[539] Yang, Q., Liu, Y., Chen, T. & Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* **10**, 1–19 (2019).

[540] Sheller, M. J. *et al.* Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports* **10**, 1–12 (2020).

[541] Zhang, P., Schmidt, D. C., White, J. & Lenz, G. Blockchain technology use cases in healthcare. In *Advances in computers*, vol. 111, 1–41 (Elsevier, 2018).

[542] Kourtis, L. C., Regele, O. B., Wright, J. M. & Jones, G. B. Digital biomarkers for alzheimer's disease: the mobile/wearable devices opportunity. *NPJ digital medicine* **2**, 1–9 (2019).

[543] Lane, N. D. *et al.* A survey of mobile phone sensing. *IEEE Communications magazine* **48**, 140–150 (2010).

[544] Plötz, T. & Guan, Y. Deep learning for human activity recognition in mobile computing. *Computer* **51**, 50–59 (2018).

[545] Guan, Y. & Plötz, T. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1**, 11 (2017).

[546] Kim, E., Helal, S. & Cook, D. Human activity recognition and pattern discovery. *IEEE pervasive computing* **9**, 48–53 (2009).